"十五"国家重点图书出版规划项目

经济学经典译丛

经济计量学

INTRODUCTION TO ECONOMETRICS

（美）詹姆斯·H.斯托克
马　克·W.沃特森　著
王庆石　主译

James H.Stock
Mark W.Watson

东北财经大学出版社

经济学经典译丛

国际贸易与竞争
企业经济学
经济学导论
宏观经济学
微观经济学
经济计量学

经济计量学

内容简介：

本教材是专为本科层次学生学习经济计量学所设计的一本初级教材，在内容上反映了当代应用经济计量学最新发展的主要内容，集中讲解那些在实践中常用的程序和检验；在教学方法上具有多样化的特点，通过多种栏目的安排有助学生理解、掌握和应用经济计量学中的核心概念；在教学方法上通过一些有趣案例的应用激发学生学习经济理论的兴趣，同时理论又与应用相结合，这一点代表了本教材与其他一些旧版本经济计量学教材的显著差别。

财经教育国际化



ISBN 7-81084-381-8
定价：56.00元



"十五"国家重点图书出版规划项目

经济学经典译丛

经济计量学

INTRODUCTION TO ECONOMETRICS

(美) 詹姆斯·H.斯托克
马克·W.沃特森 著
王庆石 主译

James H.Stock
Mark W.Watson

东北财经大学出版社

图书在版编目（CIP）数据

经济计量学／（美）斯托克（Stock, J. H.）等著；王庆石主译.—大连：东北财经大学出版社，2005.1
（“十五”国家重点图书出版规划项目·经济学经典译丛）
书名原文：Introduction to Econometrics
ISBN 7-81084-381-8

Ⅰ.经… Ⅱ.①斯… ②王… Ⅲ.计量经济学-高等学校-教材 Ⅳ.F224.0

中国版本图书馆 CIP 数据核字（2004）第013762号

辽宁省版权局著作权合同登记号：图字06-2003-261号

东北财经大学出版社出版
（大连市黑石礁尖山街217号 邮政编码 116025）
总 编 室：（0411）84710523
营 销 部：（0411）84710711
网 址：http://www.dufep.cn
读者信箱：dufep@vip.sina.com

大连海事大学印刷厂印刷 东北财经大学出版社发行

幅面尺寸：200mm×270mm 字数：682千字 印张：29 插页：1
2005年1月第1版 2005年1月第1次印刷

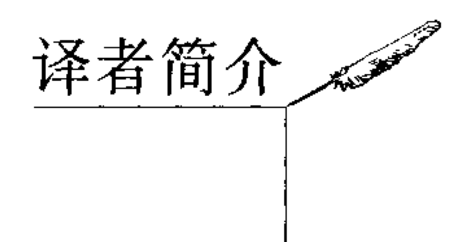
责任编辑：高 鹏 孙 越 责任校对：刘铁兰 毛 杰 尹秀英 那 欣
封面设计：张智波 版式设计：钟福建

定价：56.00元

作者简介

詹姆斯·H. 斯托克（James H. Stock）为哈佛大学经济系教授。他在加州大学伯克利分校分别获得统计学硕士学位和经济学博士学位。他的研究领域有经济计量方法、宏观经济预测、货币政策等。除《经济计量学》外，他还发表、出版了90多篇论文和其他专著。

马克·W. 沃特森（Mark W. Watson）是美国普林斯顿大学Woodrow Wilson学院经济学和公共关系学教授，也是美国经济研究局的一位副研究员。他在加利福尼亚大学圣地亚哥分校获博士学位。他的研究领域有时间序列经济计量分析、实证宏观经济学、宏观经济预测等。在上述学术领域，他发表了60多篇学术论文，曾担任美国几个有影响的杂志的编委，包括《美国经济评论》、《应用经济计量学》、《经济计量学杂志》、《商业与经济统计杂志》等。在到普林斯顿大学工作前，他曾在哈佛大学和西北大学任教。



译者简介

王庆石，男，1961年生，辽宁省辽阳县人。1983年毕业于东北财经大学统计系并获得经济学学士学位；1986年获得东北财经大学统计系统计专业经济学硕士学位；1993年获得该专业经济学博士学位。

1989年1月至1990年2月，曾在联合国人口基金开罗人口培训中心进修学习人口统计；1999年1月至1999年12月，曾以高级访问学者身份到美国纽约州立大学布法罗分校管理学院进修金融统计。现为东北财经大学数量经济学专业教授、博士生导师，兼任国际商学院院长。

中文版序

我们非常荣幸也非常高兴《经济计量学》被翻译成了中文。

经济计量学的两个主要应用，一是研究一项干预（如公共政策的变化）的因果效应，二是利用有限的历史数据做出可靠的经济预测。这两个挑战具有普遍的重要性。我们希望本书所介绍的方法、技术对中国学生也会像对我们美国学生一样很有用处。

尽管本书中的经验、例子反映的是美国的数据，但是所强调的问题却具有广泛的重要性。例如，我们在引入回归分析时所使用的主要应用案例是降低初中班级规模对学生学业表现的因果效应的估计。我们所分析的数据虽然来自美国加利福尼亚州，但基本的政策问题，即我们如何来改进青少年的教育质量，是一个跨越国界的问题。更一般地讲，本书所讲解的估计因果效应的方法已被广泛地应用到许多国家，并成功地解决了许多问题。

我们想感谢所有那些为本书在中国的翻译和出版发行做出努力的人，特别是东北财经大学出版社，以及本书的主要译者——东北财经大学数量经济系及国际商学院的王庆石教授。

詹姆斯·H. 斯托克
马克·W. 沃特森
2004. 8

前言

对学生和教师来说，经济计量学是一门充满乐趣的课程。现实的经济界、商业界和政府都是非常复杂和繁乱的领域，许多新的观念和问题需要解答。例如，对付醉酒驾车者，是通过严格的法律还是通过提高酒类的税收来限制；在股票市场上，是在价格相对较低时买进从而从中获利，还是按照股票价格随机游动理论所建议的，只需睡大觉；我们是通过缩小班级规模来改进小学的教育质量，还是每天给孩子们听十分钟的莫扎特的音乐就可以达到教育目的；等等。经济计量学可以帮助我们从纷纭复杂的问题中梳理出明确的思路，并对定量性的问题给出定量性的答案。经济计量学打开了一扇定量地认识我们这个纷纭复杂的现实世界的窗户，例如，居民、企业和政府是如何做出决策的，以及他们之间的关系如何。

本教材是专为本科层次学生学习经济计量学所设计的一门初级教材。我们的体会是，在本科的初级课程中开设经济计量学课程，通过一些有趣案例的应用激发学生学习经济理论的兴趣，同时理论又可与应用相结合。这一简单的原理代表了与其他一些旧版经济计量学教材中观点的显著差别，传统的经济计量学教材中的理论模型或理论假设与实际都是不相符的。一些学生在花费大量的时间学习经济计量学的模型或假设后，在实践中却意识到这些模型和假设是不切合实际的，这也不足为奇，所以当应用与假设不一致时他们只好重新寻求解决问题的方案。我们相信，通过具体的应用，然后再提供一些简单的与应用相符的假设，用这种方式来激发学生学习理论工具会取得更好的效果。因为理论的意义立即在应用中显现出来了，所以，这种方法使经济计量学变成一门活生生的课程。

本书的特点

本教材与其他同类教材相比，主要在三个方面有所不同。第一，我们把现实世界中的问题、数据与理论的导出结合在一起进行讲解，而且我们特别重视实证分析所取得的重要结论；第二，我们对案例的选择反映了现代的理论与实践；第三，我们提供的理论和假设与应用是相符的。我们的目的是教会学生成为经济计量学熟练的使用者，而且恰当应用本初级教材所要求掌握的数学工具。

现实世界的问题与数据

在本教材中，一些重要的现实问题往往都需要确切的数量答案，我们就是围绕这样的问题来选择每一个方法论题材的。例如，我们在讲授一元回归、多元回归和函数形式分析时，是以“估计学校的投入对学校产出的影响”（例如在小学，人数少的小型班是否比人数多的大型班考试分数更高）这一案例题材进行讲解的。我们在讲解面板数据时是通过分析“醉酒驾车法对交通死亡率的影响”这一案例来完成的。在讲授二元因变量回归



（logit 和 probit 模型）时，我们是采用“市场上家庭消费贷款中是否存在种族歧视”这一问题作为应用案例分析的。在讲授工具变量估计时，我们采用了“对香烟需求弹性的估计”这一现实案例。尽管这些例子都需要一些经济学的推理，但只要学过初级的经济学就可以懂得这些问题，许多问题即使没学过经济学照样可以明白。这样，老师在课堂上就可以专心讲授经济计量学这门课，而不是补微观经济学或宏观经济学的课。

我们非常认真地对待我们每一个实证应用的案例，我们不仅要让学生知道怎样从数据中得到信息，而且还要让学生学会自我分析，认识到实证分析的局限性。通过每一个案例的应用，我们教会学生探寻其他的分析和解决办法，并评估其结果是否稳定。在实证分析中所问的问题都是非常重要的，我们所提供的答案也是严肃的、可信的。但是，我们还是鼓励学生和老师另辟蹊径，用其他的方法分析案例中的数据。相关的内容可以在本教材的专用网站（www. aw. com/stock. watson）上查询。

关于教材的内容

经济计量学在过去的二十几年中得到了较快的发展。本教材中介绍的内容反映了当代应用经济计量学最新发展中的主要内容。一个初级课程不可能包罗万象，我们只能集中于讲解那些在实践中常用的程序和检验，例如：

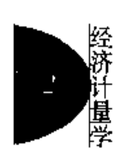
• 工具变量回归。在教材中，工具变量回归是作为处理误差项与回归因子之间相关的一般方法来介绍的，其产生有多种原因，包括联立因果关系。对一个有效工具变量的两个假设——外生性和相关性——我们给予了相同的重视。接着我们展开了讨论，工具变量从哪里来，过度识别约束的检验和弱工具变量的诊断，我们还阐释了如果这些诊断指出了问题该怎么办。

• 项目评估。越来越多的经济计量研究分析随机控制试验和准试验，又称自然试验。在第 11 章中，我们研究了这些问题，通常把它们一起视为项目评估。我们提出这种研究策略是作为遗漏变量问题、联立因果关系问题和选择问题的替代方法，我们还对使用实验数据和准实验数据的优点及缺点进行了比较和评论。

• 预测。在预测一章（第 12 章），我们使用时间序列回归方法而不是很大的联立方程结构模型分析了一元回归（自回归）和多元回归的预测问题。我们集中分析了一些简单而可靠的工具，如自回归及借助信息准则的模型选择问题，这些方法和工具在实践中都很有效。本章还特别介绍了在实践中如何处理随机趋势（单位根）、单位根检验、结构突变检验（在已知的和未知的日期）、伪样本外预测的内容，这些内容均在如何建立平稳和可靠的时间序列预测模型问题中给予了论述。

• 时间序列回归。我们明确区分了时间序列回归的两种不同应用：预测和动态因果效应估计。在使用时间序列数据进行因果推断的一章（第 13 章），对不同的估计方法（包括普通最小二乘法）何时会导致有效的因果推断、何时会导致无效的因果推断、何时应该采用带有异方差和自相关一致性标准误的 OLS 法估计动态回归，均给予了介绍。

与应用相符合的理论



尽管经济计量分析工具通过实证性应用可以使学生产生最好的理解效果，但是学生仍然需要学习足够的经济计量理论，以理解这些方法和工具的优势与局限性。我们提供了一种现代的处理方法，使理论和应用紧密结合，其涉及的数学知识仅限于高等代数。

现代的实证应用分析拥有一些共同的特征：数据集一般都较大（上百个甚至更多的

观测数据)；回归因子在重复样本中是不固定的，甚至可能是通过随机方式抽选的（或者是其他的某种随机的方法)；数据通常不服从正态分布；没有先验的理由认为误差项是同方差的（尽管经常有理由认为误差项是异方差的)。

上述特征使本教材与其他教材在对理论的建立和导出方式上有重要的差别。

• 大样本方法。因为数据集通常较大，所以，从一开始我们就使用大样本正态逼近作为假设检验和置信区间估计中抽样分布的依据。我们的体会是，讲授大样本逼近的原理花费的时间比讲授学生 t 分步、精确的 F 分步、自由度的纠正等内容花费的时间要少。大样本方法也有助于减少学生在理解他们所学的精确的分布理论并不十分重要（由于非正态误差项）这一问题上的困惑。在讲解样本均值内容时，假设检验和置信区间估计中所应用的大样本法，可直接沿续到多元回归分析、logit 和 probit 分析、工具变量估计和时间序列方法中。

• 随机抽样。因为在经济计量应用中回归因子很少是事先确定下来的，所以，我们从一开始处理各种变量（因变量和自变量）的数据，就把它们看做是随机抽样的结果。这个假定与横截面数据最初的应用是吻合的，当然也同样适用于面板数据和时间序列数据。由于采用了大样本法，就不会有其他概念上和数学上的困难。

• 异方差。应用经济计量学家经常使用异方差稳健的标准误来消除关于是否存在异方差的顾虑。在本书中，我们不是把异方差作为一个待解决的问题或例外问题来对待，相反，从一开始我们就允许异方差的存在，并直接使用异方差稳健的标准误。我们把同方差的情况作为提供 OLS 方法理论动因的一个特例。

专业化的生产者、成熟的使用者

我们希望使用此书的学生会成为实证分析方法成熟的使用者。为了实现这个目标，他们不仅要学习怎样使用这些回归分析的工具，而且还要学会如何评估他们所面对的实证分析结果的有效性。

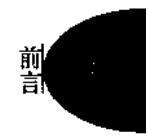
讲授如何评估一项实证研究的有效性，我们从三个方面入手。

首先，在介绍完回归分析的主要工具之后，我们就集中精力讲解第 7 章，即对一项实证研究内部有效性和外部有效性威胁因素的检验问题。这一章讨论了数据问题和将研究结论推广到其他环境中出现的问题。本章还研究了回归分析中可能存在的威胁，包括遗漏变量问题、函数形式误定问题、变量中的误差问题、变量选择问题及联立性问题，以及在实践中如何识别这些威胁的方法。

其次，我们应用这些方法评估本书中相应实证分析例子的结果。我们这样做也在考虑其他的模型设定方法，同时系统强调各种影响实证分析有效性的因素。

再次，为了成为一名成熟的使用者，学生需要像生产者一样亲自进行实验。积极的学习要胜过被动的学习，经济计量学是供学生积极学习的最为理想的一门课程。为此，本教材的网站上提供了数据集、分析软件和做各种实证分析练习的一些建议。

关于对数学背景的要求



我们的目的是使学生建立一种对现代回归分析工具的成熟的理解，不管本课程是在高水平的数学基础上讲授还是在低水平的数学基础上讲授。本书的第 1—4 部分（本部分涉及最主要的资料）对于具有初步微积分数学知识的学生来说是容易读懂的。与其他初级经济计量学教材相比，第 1—4 部分中公式较少，应用较多，当然，与数学类的本科生课

程相比公式就更少。过多的公式并不意味着是一种更为成熟的处理方法。根据我们的经验，过多的数学介绍并不一定使大多数学生理解得更深入。

这就是说，不同的学生学习的方式不同。对于数学基础较好的学生，通过较多的数学训练学习成绩可以得到明显的提高。因此第5部分包含了对经济计量学理论的介绍，这些内容对数学背景较强的学生来说较为合适。我们相信，把第5部分数学内容的章节与第1—4部分的实际资料结合在一起，可作为高年级本科生或硕士研究生经济计量学课程的教材。

本书的内容与组织

本教材分为5部分。本教材假定学生已经学过概率论与统计学的内容，尽管我们在第1章中对概率和统计的知识做了回顾。第2部分主要涉及回归分析的内容。第3部分、第4部分和第5部分在第2部分核心内容的基础上提供更深入一层的讨论议题。

第1部分

第1章是经济计量学导论，强调对定量性问题提供定量性答案的重要性。本章讨论了在统计研究中因果关系的概念，研究了经济计量学所面临的不同的数据类型。第2章和第3章分别是对概率论和统计学知识的复习，这两个章节是作为正常的章节讲授还是作为参考资料，取决于学生相关知识的背景。

第2部分

第4章介绍一个回归因子的回归分析和普通的最小二乘法（OLS法）。在第5章，学生将学到在应用多元回归时如何处理遗漏变量误差问题，进而分析当其他所有自变量保持不变时一个自变量变化的影响。在第6章，多元回归分析方法被扩展为非线性总体回归函数模型，这些模型的参数是线性的，这样可以应用OLS法进行估计。在第7章，学生们又回过头来学习如何认识回归分析的优点和局限性，以及怎样应用内部有效性和外部有效性的概念评估回归分析的结果。

第3部分

第3部分提供了回归分析方法的一些扩展。在第8章，学生将学到如何运用面板数据控制那些不随时间变化而变化的不可观测的变量。第9章介绍含有两个因变量的回归分析问题。第10章研究工具变量回归怎样用于解决在误差项与回归因子之间产生相关性的一系列问题，以及如何寻找和评估有效的工具变量。第11章引领学生进入实验或准实验（自然实验）数据的分析领域，所涉及的问题通常被称做“项目评估”。

第4部分



第4部分研究时间序列数据回归。第12章集中讲解预测方法，并介绍了各种用于分析时间序列回归的现代方法，如单位根检验和平稳性检验。第13章讨论了如何利用时间序列数据估计因果关系。第14章介绍了一些更高级的时间序列分析的工具，包括条件异方差模型。

第5部分

第5部分介绍经济计量学理论。这一部分带有一点附录性质，它补充了前面几章中省略的数学问题，但它又不仅仅是附录，它自成体系地介绍了线性回归模型中估计与推断的经济计量理论。第15章研究单个回归因子的回归分析理论，解释过程中并没有用到矩阵代数，尽管本章比其他章节确实需要有较高水平的数学推导。第16章介绍和研究了矩阵形式的多元回归模型。

本书中各章节的先后逻辑关系

因为不同的教师喜欢强调不同的资料和内容，我们写作本书时也考虑到了各种不同的教学需要。从全书的总体内容来看，第3部分、第4部分和第5部分是相对独立的，也就是说讲授这些内容不需要首先讲授先前的章节。每一章特定的先修章节在表1中给出。尽管我们发现本书讨论的问题顺序在我们自己讲授课程时很合适，但如果老师希望按不同的顺序讲解也可以，本书在章节编写时考虑到了这种需要。

表1 **第3—5部分每一章的先修章节**

章节	先修部分或章节							
	第1部分	第2部分	8.1, 8.2	10.1, 10.2	12.1—12.4	12.5—12.8	13	15
8	■	■						
9	■	■						
10.1, 10.2	■	■						
10.3—10.6	■	■	■	■				
11	■	■	■	■				
12	■	■						
13	■	■			■			
14	■	■			■	■	■	
15	■	■						
16	■	■						■

注：本表给出了讲解各章内容最低的先修章节。例如，在动态因果效应的估计一章（第13章），首先要求讲解第1部分（取决于学生的准备情况）、第2部分和第12章的第1—4节。

本书的多种用途

使用本书，可以讲解若干门课程。

标准的入门经济计量学

这门课程介绍经济计量学的一些基础知识（第1章），如果需要的话，还可以复习概率论和统计学的知识（第2章和第3章），然后进入一元回归分析、多元回归分析、函数形式分析基础和回归研究的评估（第2部分的所有内容）。这门课程还要涉及面板数据的回归分析（第8章）、含有受限因变量的回归分析（第9章），如果时间允许的话，还可



讲授工具变量回归分析（第10章）。这门课程以第11章中的实验和准实验作为结束，并重新归纳学期一开始提出的因果效应的估计问题，概述和总结回归分析的主要方法。先修课程有代数和基础统计学。

带有时间序列分析和预测应用内容的入门经济计量学

与标准的入门经济计量学相似，这门课程的内容覆盖第1部分（如果需要的话）和第2部分的全部内容。这门课程还可以提供对面板数据的简要介绍（第8.1节和第8.2节），并讲解工具变量回归（第10章，或只讲解第10.1和第10.2节），不过这些是选择性内容。然后，该课程应该讲授本书第4部分中有关预测的内容（第12章）和动态因果效应估计的内容（第13章）。如果时间允许，该课程还可以涉及一些高级的时间序列分析的内容，如条件异方差（第15.5节）。先修课程有代数和基础统计学。

应用时间序列分析与预测

本书也可以用做“应用时间序列分析与预测”课程的短期课程教材，其中回归分析是要求的先修课程。第2部分的回归分析工具也要讲授一部分，取决于学生的知识准备情况。然后，该课程的内容就可以直接转至第4部分，从预测（第12章）、动态因果效应的估计（第13章）到时间序列分析的高级议题（第14章），包括向量自回归和条件异方差。这门课程一个非常重要的部分是学生亲手做预测作业，这些作业可在本书的专门网站上得到。先修课程有代数和基础经济计量学或同类课程。

经济计量学理论入门

本书也适合于用做具有较好数学基础的高年级本科生的经济计量学理论的教材，或者用做硕士研究生的经济计量学教材。如果有必要，该课程可以先复习概率论和统计学的理论（第1部分）。该课程可以采用非数学的方式即以应用为基础的方式介绍回归分析（第2部分），接着应该介绍经济计量学的理论，即第15章和第16章，然后再讲解带有受限因变量的回归分析（第9章）和极大似然估计法（附录9.2），最后应转入讲解工具变量回归分析（第10章）、时间序列分析（第12章），以及应用时间序列数据和普通最小二乘法（第13章和第16.6节）讲解因果效应估计，当然最后这部分内容可选可不选。先修课程有微积分和基础统计学。第16章还要求学生具有一定的矩阵代数的知识。

教学方法上的安排

本教材在教学方法上具有多样化的特点，意在帮助学生理解、掌握和应用经济计量学中的核心概念。“章节介绍”部分提供了一个符合现实的介绍和我们编排的动因，还给出了描绘各章讨论内容的路径图。“关键性词汇”以清晰的结构定义了各章的主要词汇。在固定间距的“重要概念”框中简要地归纳了一些重要的概念。“一般兴趣”框则对相关话题和一些现实中运用所讲概念和方法所做的研究成果提供了一些有趣的讨论。“总结”部分则对复习每一章主要内容的要点是很有帮助的。在“复习概念”部分所提出的问题，是检查学生对所学核心内容的理解能力，“练习”部分则给出了一些运用本章介绍的概念和技术解决实际问题的练习题。在教材的最后，“参考文献”部分列出了供进一步阅读的资料，“附录”则提供了统计表，“术语表”（词汇）简洁地定义了本书中出现的所有关



键性术语。

致谢

许多人为此书的完成做出过贡献。首先，我们要特别感谢我们在哈佛大学和普林斯顿大学的同事们，他们在课堂教学中曾使用过本书的初稿。在哈佛大学肯尼迪政府学院，Suzanne Cooper 对几个草稿都提出过中肯的建议和详细的评论。她作为与本书作者之一斯托克（Stock）的合作教师，在本书作为肯尼迪政府学院硕士研究生的必修课教材的讲授期间，曾帮助审查过本书的大部分资料。其次，我们要感谢肯尼迪政府学院的另外两位同事 Alberto Abadie 和 Sue Dynarski，谢谢他们对准实验和项目评估领域所做的耐心解释，以及他们对本书早期初稿的详细评论。在普林斯顿大学，Eli Tamer 讲授过本书早期的一个初稿，并对本书出版前的一个初稿提供过很有帮助的评论。

我们还要感谢那些在经济计量学领域就本书的核心内容与我们做过许多交流并提过许多有价值的建议的朋友和同事。Bruce Hansen（威斯康星大学麦迪逊学院）和 Bo Honore（普林斯顿大学）对本书早期的大纲和第 2 部分核心资料的初稿提供过很有帮助的反馈。Joshua Angrist（麻省理工学院）和 Duido Imbens（加州大学伯克利分校）对项目评估资料在本书中的处理方式为我们提供过深思熟虑的建议。本书中对时间序列资料的表现方式的论述得益于与 Yacine Ait-Sahalia（普林斯顿大学），Graham Elliott（加州大学圣地亚哥分校），Andrew Harvey（剑桥大学）以及 Christopher Sims（普林斯顿大学）的讨论。最后，很多人就他们所熟悉的领域对本书手稿的部分内容提出过有价值的建议。他们是：Don Andrews（耶鲁大学），John Bound（密歇根大学），Gregory Chow（普林斯顿大学），Thomas Downes（Tufts 大学），David Druckker（Stata 有限公司），Jean Baldwin Grossman（普林斯顿大学），Eric Hanushek（斯坦福大学胡佛学院），James Heckman（芝加哥大学），Han Hong（普林斯顿大学），Caroline Hoxby（哈佛大学），Alan Krueger（普林斯顿大学），Steven Levitt（芝加哥大学），Richard Light（哈佛大学），David Neumark（密歇根州立大学），Joseph Newhouse（哈佛大学），Pierre Perron（波士顿大学），Kenneth Warner（密歇根大学）和 Richard Zeckhauser（哈佛大学）。

还有许多人非常慷慨地给我们提供了数据。加利福尼亚州考试分数数据是在加利福尼亚州教育局标准与评估部的 Les Axelrod 的帮助下构建完成的。我们感谢马萨诸塞州教育局学生评估服务部的 Charlie DePascale，他对我们获得马萨诸塞州考试分数数据集提供过帮助。Christopher Ruhm（北卡罗来纳大学格林保罗分校）很亲切大方地向我们提供了他的关于醉酒驾车法和交通事故死亡方面的数据集。波士顿联邦储备银行的研究部把在抵押贷款种族歧视方面的数据合成起来提供给我们，对此我们非常感谢。我们特别要感谢 Geoffrey Tootell，他向我们提供了第 9 章中使用的最新的数据，同时特别要感谢 Lynn Browne，他把这些数据集的政策性含义解释给我们。我们感谢 Jonathan Gruber（麻省理工学院），他让我们分享了他的关于香烟销售方面的数据（即第 10 章中我们分析的数据）。感谢 Alan Krueger（普林斯顿大学），他帮助我们获得了田纳西州 STAR 数据集，即我们在第 11 章中分析的数据。



此外，我们还要感谢所有那些为 Addison-Wesley 出版公司审查过本书的初稿并提出过许多详尽的、深思熟虑的、有建设性评论的人。他们是：Michael Abbott（加拿大女王学院），Alok Bohara（新墨西哥大学），Richard J. Agnello（德拉华大学），Chi-Young Choi

（新汉普郡大学），Clopper Almon（马里兰大学），Dennis Coates（马里兰大学巴尔的摩县分校），Joshua Angrist（麻省理工学院），Tim Conley（芝加哥大学商学院），Swarnjit S. Arora（威斯康星大学密尔沃基分校），Douglas Dalenberg（蒙大拿大学），Christophger F. Baum（波士顿学院），Antony Davies（迪尤肯大学），McKinley L. Blackburn（南卡罗来纳大学），Joanne M. Doyle（詹姆斯・麦迪逊大学），Divid Eaton（默里州立大学），Mico Mrkaic（杜克大学），Adrian R. Fleissig（加利福尼亚州立大学富勒顿分校），Serena Ng（约翰・霍普金斯大学），Rae Jean B. Goodman（美国海军军官学校），Jan Ondriech（雪城大学），Bruce E. Hansen（威斯康星大学麦迪逊分院），Pierre Perron（波士顿大学），Peter Reinhard Hansen（布朗大学），Robert Phillips（乔治・华盛顿大学），Ian T. Henry（澳大利亚墨尔本大学），Simran Sahi（明尼苏达大学），Marc Henry（哥伦比亚大学），Sunil Sapra（加利福尼亚州立大学洛杉矶分院），William Horrace（亚利桑纳大学），Frank Schortfheide（宾夕法尼亚大学），Oscar Jorda（加利福尼亚大学戴维斯分校），Leslie S. Stratton（弗吉尼亚联邦大学），Frederick L. Joutz（乔治・华盛顿大学），Jane Sung（杜鲁门州立大学），Elia Kacapyr（以色佳学院），Christopher Taber（西北大学），Manfred W. Keil（科莱蒙特・麦肯纳学院），Petra Todd（宾夕法尼亚大学），Eugene Kroch（威拉诺瓦大学），John Veitch（旧金山大学），Gary Krueger（麦卡莱斯特学院），Edward J. Vytlacil（斯坦福大学），Kajal Lahiri（纽约州立大学阿尔巴尼分院），M. Daniel Westbrook（乔治敦大学），Daniel Lee（夕本斯堡大学），Tiemen Woutersen（西安大略大学），Tung Liu（保尔州立大学），Phanindra V. Wunnava（米德伯里学院），Ken Matwiczak（德克萨斯大学 LBJ 公共事务学院奥斯汀分院），Zhenhui Xu（乔治亚州立大学），KimMarie McGoldrick（里士满大学），Yong Yin（纽约州立大学布法罗分校），Robert McNown（科罗拉多大学博德分校），Jiangfeng Zhang（加利福尼亚大学伯克利分校），H. Naci Mocan（科罗拉多大学丹佛分校），John Xu Zheng（德克萨斯大学奥斯汀分校）。

我们还要感谢细心校对本书书稿的几个人。Kerry Griffin 和 Yair Listokin 阅读了整个手稿；Andrew Fraker，Ori Heffretz，Amber Henry，Hong Li，Alessandro Tarozzi 以及 Matt Watson 也审查了几章的内容。

我们的工作得益于杰出的编辑 Jane Tufts 的帮助，他的智慧和勤奋以及严谨的工作作风使本书增色不少。Addison-Wesley 公司向我们提供了一流的支持，从最优秀的编辑 Sylvia Mallory，到整个出版团队。Jane 和 Sylvia 耐心地教授我们很多关于写作、组织及表现的技巧，他们的努力在本书的每一页中都可以见到。我们还想把我们的感谢献给卓越的 Addison-Wesley 公司的整个团队，本书出版中的每一个过程都凝聚了他们的奉献：Adrienne D. Ambrosio（营销经理）、Melissa Honig（高级媒体制作师）、Regina Kolenda（高级设计师）、Katherine Watson（制作总监），特别是 Denise Clinton（总编）。

最后，我们要致谢我们的家人，谢谢他们在本项目进行过程中所表现出的耐心。写作本书花费了很长的时间，对我们的家人来说这真是无休止的工作，他们承受了许多痛苦。对他们的帮助和支持我们致以深深的谢意。

目　录

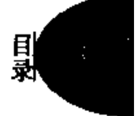
第1部分

引言与相关知识的复习

第 2 部分

回归分析基础







第 5 部分

回归分析的经济计量理论

第 1 部分

引言与相关知识的复习

第1部分

第1章 经济问题与数据

如果问六个经济计量学家这样的问题:什么是经济计量学?你可能会得到六个不同的答案。第一个人可能会告诉你经济计量学是一门检验经济理论的学科;第二个人可能会告诉你经济计量学是用来预测诸如公司的销售额、经济的总体增长率或股票价格等经济变量未来值的一整套工具;第三个人可能会说经济计量学是用数理经济模型拟合现实世界经济数据的过程;第四个人可能会告诉你,在政府和商业部门,经济计量学是运用历史数据做出数值化或定量化政策建议的一门艺术和科学。

实际上,这些答案都是正确的。广义上来说,经济计量学是一门运用经济理论和统计技术来分析经济数据的科学和艺术。经济计量学方法在经济学的许多分支中都有广泛的应用,包括金融学、劳动经济学、宏观经济学、微观经济学、市场营销学和经济政策制定等。此外,经济计量学方法在政治学和社会学等其他社会科学领域中也有较多的应用。

本书将向你介绍经济计量学家们常用的一整套经济计量的核心方法。我们将运用这些方法解答来自商业界和政府政策制定中各种具体的和定量的问题。本章中列举了四个此类问题,并用最普通的术语探讨了解决这些问题的经济计量方法。为了回答这些以及其他的定量性经济问题,经济计量专家需要取得数据,因此本章以探讨各种数据类型的调查方法作为本章的结束。

1.1 我们所研究的经济问题

经济界、商业界和政府部门中的许多决策都依赖于我们对身边世界中各种变量之间关系的理解。这些决策要求对定量的问题给出定量的答案。

本书选择了当前经济问题中四个定量性的经济问题。这四个问题涉及教育政策、抵押借贷的种族歧视、香烟消费和宏观经济预测。



1.1.1 问题1:减小班级规模会改进小学的教育水平吗

对美国公立教育系统改革的提议引发了激烈的争论。许多提议关心低年级学生即小学

1.1.3 问题3:香烟税在减少香烟消费方面起多大作用

吸烟问题是世界范围内关注的公共健康问题。许多吸烟的成本,比如因照顾由吸烟致病的患者所发生的医疗费用,以及那些不好量化的因间接吸烟引致的疾病的成本,都是由社会其他成员来承担的。因为这些费用要由社会大众而不是吸烟者来承担,所以,政府此时要进行干预以减少香烟消费。减少香烟消费最灵活的方法之一就是提高香烟税。

经济学基本原理认为,如果香烟价格上升,香烟消费就会减少。但是会减少多少呢?如果销售价格上涨了1%,香烟的销售量会下降百分之多少呢?由价格增加1%所导致的需求量变化的百分比就是需求的价格弹性。如果我们要通过增加税收来减少一定数量的香烟消费,比如说20%,那么,我们需要知道价格弹性才能计算出为使消费量下降20%需要上涨的价格幅度。但香烟的需求价格弹性是多少呢?

尽管经济学理论为我们提供了回答这个问题的概念,但它没有告诉我们需求价格弹性的数值。为了了解这个弹性值,我们必须分析吸烟者和潜在吸烟者的行为数据。换句话说,我们需要分析香烟的消费数据和价格数据。

我们分析的数据是20世纪80年代和90年代美国各州的香烟销售量、价格、税收和个人收入。这些数据表明,具有低香烟税进而香烟价格较低的州的吸烟率高,而具有高香烟税进而香烟价格较高的州的吸烟率就低。不过,这些数据的分析是复杂的,因为因果关系可能向两个方向走:低税收会导致高需求。但是,如果一个州有很多的吸烟者,那么该州的政客就可能会尽力保持低香烟税以取悦他们的吸烟选民。在第10章中,我们研究了处理这种"联立因果关系"的经济计量方法,并运用这些方法估计了香烟需求的价格弹性。

1.1.4 问题4:明年的通货膨胀率将会是多少

人们似乎总是私下里想要对未来进行预测。例如,一个正在考虑投资于新设备的公司明年的销售量会是多少?下个月股票市场会上涨吗?如果会,那么上涨多少呢?明年的城市税收入能否够支付预算内的城市服务费用呢?下周的宏观经济学课程考试会集中在外部性问题上,还是在垄断性问题上?这个周六是否是个适合郊游的好天气?

宏观经济学家和金融经济学家们对未来特别感兴趣的一个方面是明年总体的价格通货膨胀水平。金融专家可能会建议客户在给定的利率下是贷款还是还贷,取决于她对来年通货膨胀率的最佳猜测。像在华盛顿的联邦储备委员会、德国法兰克福的欧洲中央银行这样的中央银行任职的经济学家们负有维持价格通货膨胀水平稳定的责任,因此,他们制定利率的决策依赖于他们对明年通货膨胀率的展望。如果他们认为明年的通货膨胀率要上涨某个百分点,那么他们可能会提高利率,其提高的幅度比通货膨胀的幅度还要高,以减缓经济过热发展带来的风险。如果他们猜测错误,那么他们也会冒风险,即可能导致经济不必要的衰退,或者使通货膨胀率涨得更高。

那些依赖于精确数值预测结果的专业经济学家,是用经济计量模型做出预测的。预测者的工作就是利用过去来预测未来,经济计量学家则是通过运用经济理论和统计技术来量化历史数据之间的关系进而对未来做出预测的。

我们用来预测通货膨胀的数据是美国的通货膨胀率和失业率。宏观经济数据中一个重要的经验关系是著名的"菲利普斯曲线",该曲线指出,当前的低失业率将伴随着来年通货膨胀率的上涨。在第12章,我们所建立并进行评价的通货膨胀预测,其结果之一就是以菲利普斯曲线为基础做出的。





1.1.5 定量的问题需要给出定量的答案

这四个问题中的每一个都需要数值答案。经济理论提供了关于这些答案的线索。比如,价格上升时,香烟消费量应该下降,但是确切的数值必须要经过实证分析才能得到,即分析实际数据。因为我们要用数据来回答定量的问题,所以,我们的答案总有一定程度的不确定性,即不同的数据集会产生不同的数值答案。因此,分析的概念框架既要提供所关心问题的数值答案,又要提供答案准确性的测度。

本书使用的概念框架是多元回归模型,多元回归模型是经济计量学的支柱。本书第2部分所介绍的这种模型,提供了一种在保持其他因素不变的条件下量化一个变量的变化如何影响另一个变量的数学方法。例如,在保持学生特征(如家庭收入)不变的条件下(这些特征学区管理者无法控制),班级规模变化对考试成绩有什么影响?在保持诸如还贷能力等其他因素不变的条件下,申请人的种族对其获得住房抵押贷款的机会会有什么影响?在保持吸烟者和潜在吸烟者收入不变的条件下,香烟价格提高1%对香烟消费会有什么影响?多元回归模型及其扩展模型提供了用数据回答这些问题,以及量化与这些问题答案相关的不确定性的框架。

1.2 因果效应和理想化实验

和经济计量学中遇到的许多其他问题一样,1.1节中的前三个问题都涉及变量间的因果关系。一般地说,如果某种结果是某种行为的直接后果或结论,那么,就称该行为是引致该结果的原因。例如,碰到热火炉会导致烫伤;喝水能解渴;轮胎充气就会膨胀;给西红柿施肥会使它们结出更多的果实。因果关系是指,某一特定行为(如施肥)导致了某一特定的可测度的结果(如更多的西红柿)。

1.2.1 因果效应的估计

我们怎样才能精确地测度一定的施肥量,比方说每平方米100克肥料,对西红柿产量的因果效应(用千克测度的)?

测度这种因果效应的方法之一就是进行一项实验。在这个实验中,园艺研究人员在多块地中种植了西红柿,给每块地以同样的管理,只有一点除外,即一些地块每平方米施肥100克,而另一些地块则不施肥。此外,某块地施肥与否是由计算机随机决定的,这样确保了地块之间的任何其他差异都和施肥与否无关。在收获期,园艺家计算出每块地的产出,施肥的地块和没施肥的地块平均每平方米产量之差就是施肥对西红柿产量的因果效应。

这是随机化控制实验(randomized controlled experiment)的一个例子。说它是控制实验,是因为存在一个控制组(control group)(地块),这个组中没有给予施肥处理;还有一个处理组(treatment group),这个组中给予了施肥处理(每平方米100克的肥料)。我们说实验是随机化的,是指处理是按随机原则分配的。这种对处理的随机分配排除了诸如某块地的日照情况怎样和它是否得到施肥两组之间系统关系的可能性,因此在控制组和处理组之间惟一的系统差异就是施肥处理。如果该实验能够在足够大的规模上进行,那么,我们就将会得到一个因果效应的估计值,即处理(每平方米施100克肥料)对所感兴趣的结果(西红柿产量)的因果效应估计值。

本书中,因果效应(causal effect)被定义为一个给定的行为或处理对某一结果的影响,





就像在一个理想的随机化控制实验中所测度的那样。在这样的实验中,处理组与控制组之间的结果产生差异的惟一的系统原因就是处理本身。

我们可以设想用一个理想的随机化控制实验来回答 1.1 节中提出的前三个问题。例如,为了研究班级规模问题,我们可以假设将不同班级规模的"处理"随机地分配到不同的学生群组。如果设计并执行了这个实验,使学生群组之间惟一的系统差异是他们的班级规模,那么理论上在保持其他条件不变的情况下,这个实验将会估计出班级规模对考试成绩的影响。

理想化的随机化控制实验是一个非常有用的概念,因为它给出了因果效应的定义。然而,实际上进行理想化的实验是不可能的。事实上,经济计量学中实验是很少见的,因为这些实验或者常常不符合道德标准,或者不可能得到很满意的执行,或者代价非常高昂。尽管如此,理想的随机化控制实验的概念确实为使用实际数据进行因果效应的经济计量分析提供了一个理论基准。

1.2.2 预测与因果关系

尽管 1.1 节中的前三个问题涉及因果关系,但第四个问题,即预测通货膨胀,却不涉及因果关系。为了做出一个好的预测,并不一定需要知道因果关系。要"预测"是否正在下雨,一个好的办法是观察行人有没有打雨伞,但打雨伞的行为并不是引起下雨的原因。

尽管预测并不需要涉及因果关系,但宏观经济理论所建立起来的模型和关系,对预测通货膨胀却是很有用的。我们在第 12 章将会看到,多元回归分析允许我们把经济理论所揭示的历史关系量化出来,允许我们检验这些关系是否随着时间的推移平稳地变化,允许我们对未来进行定量的预测,允许我们对这些预测结果的精度进行评价。

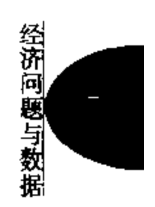
1.3 数据:来源与类型

经济计量学中的数据主要有两个来源:对现实世界的实验或非实验观测。本教材既分析实验数据集,也分析非实验数据集。

1.3.1 实验数据与观测数据

实验数据(experimental data)来源于实验,这些实验或者是设计用来评价一项处理或一个政策,或者是用来检验一项因果效应。例如,在 20 世纪 80 年代,田纳西州资助了一项研究班级规模对学习效果影响的大型随机化控制实验。在该项实验中(第 11 章我们将分析该实验的数据),成千上万的学生被随机地分配到不同规模的班级中,研究人员连续观察几年,每年都进行标准化考试,记录考试的结果。

田纳西州的班级规模实验耗费了大量的资金,并且要求众多的管理者、家长和教师进行了连续好几年的合作。因为现实生活中以人为研究对象的实验很难管理和控制,所以,相对于理想随机化控制实验它们还存在很多不足之处。此外,在某些环境下,实验不仅昂贵和难于管理,而且还可能不符合道德标准。(比如,向随机选定的十几岁的青少年提供廉价的香烟以了解他们会买多少,这项实验是符合道德标准的吗?)由于这些财务经费的原因、实际可行性的原因以及道德的原因等,经济学中很少进行实验。相反,大多数经济数据是通过观测现实世界的行为得到的。



通过观测实验之外的现实世界中的实际行为所得到的数据,称为观测数据

(observational data)。观测数据是利用抽样调查(如对消费者的电话调查)和行政管理记录(如贷款机构保存的关于抵押贷款申请者的历史记录)搜集的。

观测数据对试图估计因果效应的经济计量学提出了挑战,然而,经济计量学的工具就是处理这些挑战的。在现实生活中,"处理"的水平(如西红柿案例中的施肥量、班级规模案例中的学生—教师比)并不是随机分配的,因此很难将"处理"的影响从其他的相关因素中分离出来。经济计量学的大部分内容,也即本书的大部分内容,都是致力于研究在用现实数据估计因果效应时所遇到的挑战及其解决方法的。

无论数据是实验数据集还是观测数据集,一般都有三种主要类型:截面数据、时间序列数据、面板数据。在本书中这三类数据都会遇到。

1.3.2 截面数据

不同实体(如工人、消费者、企业、政府单位等)在某一单一时期的数据称为**截面数据**(cross-sectional data)。例如,加利福尼亚州各学区考试成绩的数据就是截面数据。因为这些数据是420个实体(学区)在单一时期(1998年)的数据。一般地说,用n来表示我们所观测到的实体个数,因此在加利福尼亚州的数据集合中,$n=420$。

在加利福尼亚州的考试成绩数据集中,每一个学区都包含有对不同变量的测度值。表1—1列出了其中的部分数据。每一行列出了不同学区的数据。例如,第一个学区(学区1)的平均考试成绩是690.8,这是该区1998年标准化考试(斯坦福达标考试)中所有五年级学生的数学和自然科学的平均成绩;该地区的学生—教师比是17.89,即学区1的学生人数被学区1的任课教师人数除得到17.89。第一个学区每个学生的平均费用是6 385美元。该学区中仍在学习英语的学生的百分比(即把英语作为第二语言且对英语还不精通的学生的百分比)是0%。

表1—1　　1998年加利福尼亚州学区的考试成绩和其他变量的部分观测值

观测值(地区)编号	地区平均考试成绩(五年级)	学生—教师比	每个学生的费用(美元)	学习英语的学生比例(%)
1	690.8	17.89	6 385	0.0
2	661.2	21.52	5 099	4.6
3	643.6	18.70	5 502	30.0
4	647.7	17.36	7 102	0.0
5	640.8	18.67	5 236	13.9
⋮	⋮	⋮	⋮	⋮
418	645.0	21.89	4 403	24.3
419	672.2	20.20	4 776	3.0
420	655.8	19.04	5 993	5.0

注:加利福尼亚州考试成绩数据集在附录4.1中介绍。



余下各行代表其他学区的数据。各行的顺序是任意排列的,学区的标号(称为**观测序号**(observation number))也是任意分配的数字,以便于排序。从表1—1中可以看出,列出的所有变量值变化都相当大。

利用截面数据,通过研究在单一时期内个人、企业或其他经济实体之间的差异,我们可



表1—3　按照州和年份表示的美国各州香烟销售量、价格和税率的部分观测值(1985—1995)

观测值编号	州　名	年份(年)	香烟销售量(包/人)	每包平均价格(美元)(含税)	总税额(美元)(香烟特许权税+销售税)
1	亚拉巴马州	1985	116.5	1.022	0.333
2	阿肯色州	1985	128.5	1.015	0.370
3	亚利桑那州	1985	104.5	1.086	0.362
⋮	⋮	⋮	⋮	⋮	⋮
47	西弗吉尼亚州	1985	112.8	1.089	0.382
48	怀俄明州	1985	129.4	0.935	0.240
49	亚拉巴马州	1986	117.2	1.080	0.334
⋮	⋮	⋮	⋮	⋮	⋮
96	怀俄明州	1986	127.8	1.007	0.240
97	亚拉巴马州	1987	115.8	1.135	0.335
⋮	⋮	⋮	⋮	⋮	⋮
528	怀俄明州	1995	112.2	1.585	0.360

注:香烟消费的数据集在附录10.1中介绍。

表1—3中列出了香烟消费数据集中的部分数据。前48行的观测值从亚拉巴马州到怀俄明州按字母排序列出了1985年各州的数据。接下来的48行的观测值列出了1986年的数据,依此类推,直到1995年。例如,在1985年,阿肯色州的香烟销售量是每人128.5包(该州1985年销售的香烟总包数除以该州1985年的总人口数得到128.5)。阿肯色州1985年每包烟含税的平均价格是1.015美元,其中有37美分被缴纳了联邦税、州税和地方税。

根据面板数据,可以了解数据集中多个不同实体的经济关系,以及每个实体随时间变化的演变关系。

截面数据、时间序列数据和面板数据的定义在重要概念1.1中进行总结。

重要概念

截面数据、时间序列数据和面板数据

■截面数据由多个实体在单一时期的观测值组成。

■时间序列数据由单一实体在多个时期的观测值组成。

■面板数据(也称纵向数据)由多个实体在两个或两个以上时期的观测值组成。

总结

1. 许多商业和经济中的决策都需要知道一个变量的变化如何影响另一个变量变化的定量估计值。

2. 从理论上说,估计一个因果效应的方法是在一个理想的随机化控制实验下进行的,但在经济应用中进行这样的实验通常是不符合道德标准的、不实际的,或是过于昂贵的。

3. 经济计量学提供了利用观测(非实验的)数据或来自于现实世界的不完备的实验数

据估计因果效应的工具。

4. 截面数据是通过在单一时点上观测多个实体得到的;时间序列数据是通过在多个时点上观测单一实体得到的;而面板数据是通过观测多个实体得到的,其中每个实体都在多个时点上被观测。

重要术语

随机化控制实验　控制组　处理组　因果效应　实验数据　观测数据　截面数据　观测序号　时间序列数据　面板数据　纵向数据

复习概念

1.1　设计一个假设的、理想的随机化控制实验来研究学习时间对微观经济学课程考试成绩的影响,并列出实际上执行这个实验的障碍。

1.2　设计一个假设的、理想的随机化控制实验来研究系安全带对高速公路交通事故死亡率的影响,并列出实际上执行这个实验的障碍。

1.3　如果要你研究一个制造厂职工的培训时间(用每个职工每周的小时数测量)与职工的生产率(每个员工每小时的产出量)之间的关系,请叙述:

a. 测度这个因果效应的一个理想的随机化控制实验;

b. 你用来研究这个因果效应的截面观测数据集;

c. 你用来研究这个因果效应的时间序列观测数据集;

d. 你用来研究这个因果效应的面板观测数据集。

2.1.1 概率、样本空间和随机变量

概率和结果 你遇到的下一个陌生人的性别，你的考试成绩，你在写学期论文时电脑死机的次数，这些都含有偶然因素或随机性。在每一个例子中，都存在一定的未知性，但这种未知性最终会被揭晓。

一个随机过程所产生的相互排斥的可能后果被称为结果(outcomes)。例如，你的电脑可能从不死机，可能死机1次，也可能死机两次，如此等等。这些结果中只有一个会真正发生(结果是相互排斥的)，而且这些结果的发生的可能性并不需要是等同的。

结果的概率(probability)是指这个结果长期发生次数的比率。如果在你写一篇学期论文的时候，电脑不死机的概率是80%，那么在写许多篇学期论文的过程中，你会完成80%的写作任务，同时电脑不死机。

样本空间与事件 所有可能结果的集合被称为样本空间(sample space)。事件(event)是样本空间的一个子集，即事件是一个或多个结果的集合。“我的电脑死机不超过1次”这个事件是由两个结果组成的集合：“不死机”和“死机1次”。

随机变量 随机变量是一个随机结果的一系列数值表示。当你写一篇学期论文时，你的电脑死机的次数是随机的，并取一个特定的数值，因此它就是一个随机变量。

一些随机变量是离散的，一些随机变量是连续的。顾名思义，离散型随机变量(discrete random variable)只能取离散的数值，如0,1,2,…，而连续型随机变量(continuous random variable)则取可能值的连续区间。

2.1.2 离散型随机变量的概率分布

概率分布 离散型随机变量的概率分布(probability distribution)就是变量的所有可能值和每个值发生的概率的列表。这些概率之和等于1。

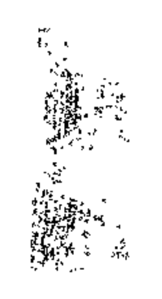
例如，用M表示在你写学期论文时电脑死机的次数。随机变量M的概率分布就是每个可能结果的概率的列表；$M=0$的概率是电脑不死机的概率，表示为$\Pr(M=0)$；$\Pr(M=1)$是电脑死机1次的概率；依此类推。表2—1的第2行给出M概率分布的一个例子。在这个分布中，如果你的电脑死机4次，你将放弃使用电脑而改用手写论文。根据这个分布，不死机的概率为80%；死机1次的概率为10%；死机2次、3次或4次的概率分别为6%，3%和1%。这些概率之和等于100%。这个概率分布绘制在图2—1中。

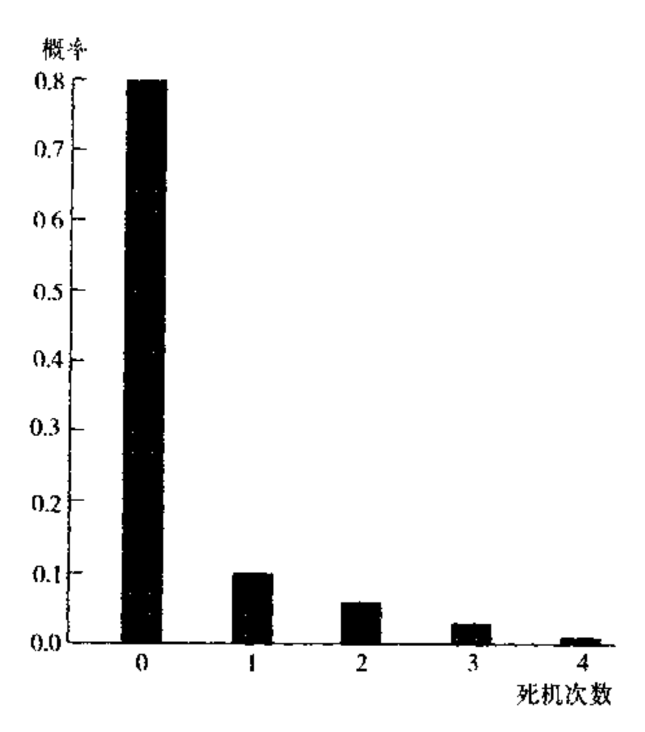
表2—1 你的电脑死机M次的概率

	结果(死机次数)				
	0	1	2	3	4
概率分布	0.80	0.10	0.06	0.03	0.01
累积概率分布	0.80	0.90	0.96	0.99	1.00

事件的概率。事件的概率可以从概率分布中计算。例如，死机1次或2次事件的概率是表中对应结果的概率和，即$\Pr(M=1\text{或}M=2)=\Pr(M=1)+\Pr(M=2)=0.10+0.06=$







注：每个柱形的高度是所标明的电脑死机次数的概率。第一个柱形的高度是0.80，因此，电脑死机0次的概率是80%；第二个柱形的高度是0.1，因此，电脑死机1次的概率是10%；其他柱形的含义依此类推。

图2—1　电脑死机次数的概率分布

0.16，或表示为16%。

累积概率分布　累积概率分布（cumulative probability distribution）是随机变量小于或等于某个特定值的概率。表2—1的最后一行给出了随机变量M的累积概率分布。例如，至多死机1次的概率$\Pr(M\leqslant 1)$是90%，它是不死机的概率（80%）和死机1次的概率（10%）之和。

累积概率分布也常被称为累积分布函数（（cumulative distribution function）c. d. f.），或累积分布（cumulative distribution）。

贝努里分布　离散型随机变量的一个重要的特殊情形就是随机变量是二元的，即结果是0或1。二元随机变量被称为贝努里随机变量（Bernoulli variable）（为纪念17世纪瑞士数学家和科学家Jacob Bernoulli），它的概率分布被称为贝努里分布（Bernoulli distribution）。

例如，假设G是你遇到的下一个陌生人的性别，其中$G=0$表示这个人是男性，$G=1$表示这个人是女性。因而，G的结果和对应的概率为：

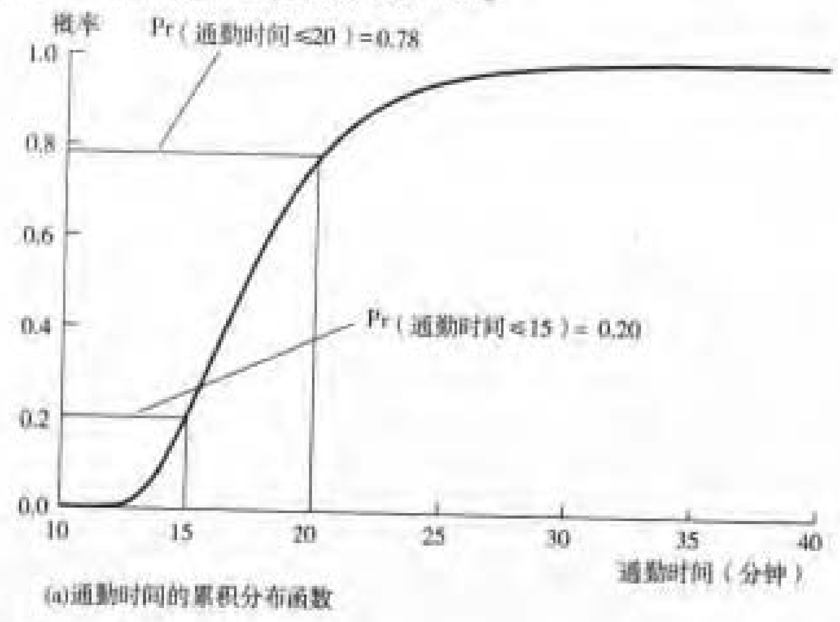
$$G=\begin{cases}1 & \text{概论为 } p \\ 0 & \text{概率为 } 1-p\end{cases} \tag{2.1}$$

其中，p是你遇到的下一个陌生人为女性的概率。公式（2.1）中的概率分布就是贝努里分布。

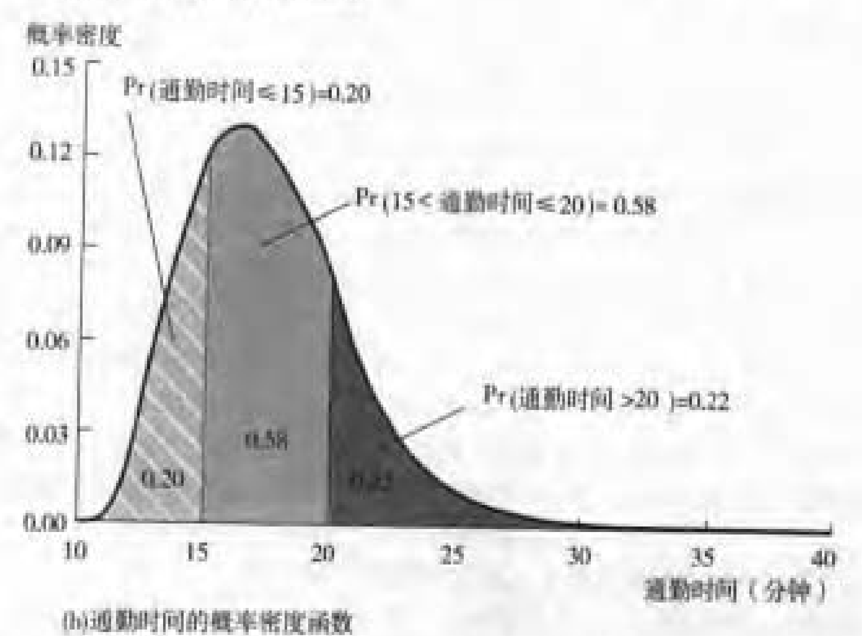
2.1.3　连续型随机变量的概率分布

累积概率分布　连续型随机变量的累积概率分布的定义与离散型随机变量的累积概率分布类似。也就是说，连续型随机变量的累积概率分布是随机变量小于或等于某个特定值的概率。

例如,考虑一个从家开车到学校的学生。这个学生的通勤时间可以取某值的一个连续的区间,因为它依赖于诸如天气和交通状况的随机因素,自然地将其看做是连续型随机变量。图2—2(a)描绘了一个通勤时间假定的累积概率分布。例如,通勤花费的时间少于15分钟的概率是20%,不超过20分钟的概率是78%。



(a)通勤时间的累积分布函数



(b)通勤时间的概率密度函数

注:图2—2(a)描绘了通勤时间的累积概率分布,通勤时间少于15分钟的概率为20%,通勤时间多于20%的概率为78%。图2—2(b)描绘了通勤时间的概率密度函数。概率由累积密度函数的面积给出,通勤时间介于15分钟和20分钟之间的概率为58%,概率面积为相应曲线下的部分。

图2—2 通勤时间的累积分布与概率密度函数

概率密度函数。由于连续型随机变量能够取所有可能值的一个连续的区间,因此,用于列出离散型随机变量每个可能值的概率所使用的概率分布的方法就不适合连续型随机变量了。相反,连续型随机变量是用概率密度函数(Probability density function)来刻画的。在概率密度函数下,任意两点间的面积就是该随机变量落在那两个点之间的概率。概率密度



函数还被称为 p. d. f. 、密度函数(density function),或简称为密度(density)。

图2—2(b)描绘了与图2—2(a)累积概率分布相对应的通勤时间的概率密度函数。在15~20分钟之间的p. d. f. 下的面积给出了通勤时间在15至20分钟之间的概率,它是0.58或58%。同样,这个概率值也可以通过图2—2(a)中的累积分布图反映出来,即用通勤时间少于20分钟的概率(78%)减去通勤时间少于15分钟的概率(20%)所得之差。因此,概率密度函数和累积概率分布是以不同的形式表达相同的信息。

2.2 期望值、均值和方差

2.2.1 随机变量的期望值

期望值。随机变量 Y 的**期望值**(expected value)是多次重复试验或发生的过程中随机变量的长期平均值,表示为 $E(Y)$。计算离散型随机变量的期望值,是将该随机变量的可能结果加权平均,这里权数就是对应的结果的概率。变量 Y 的期望值也被称为 Y 的**期望**(expection),或 Y 的**均值**(mean),表示为 μ_Y。

例如,假设你以10%的利率贷款100美元给你的一位朋友。如果贷款得到偿还,你就可以得到110美元(100美元本金加上10美元利息),但如果你的朋友有1%的违约风险,你就会什么也得不到。因此,你得到的偿还数量是个随机变量,你有0.99的概率得到110美元,有0.01的概率得到0美元。在许多这样的借贷中,你得到偿付110美元的机会有99%,但什么都得不到的机会有1%,因此平均来看,你将被偿付110×0.99+0×0.01=108.90(美元)。这样,你的偿付期望值(或"平均偿付")是108.90美元。

另一个例子,考虑表2—1中给定的概率分布的电脑死机次数 M。M 的期望值是在许多学期论文写作期间的平均死机次数,以给定大小死机次数的发生频率为权数,可以得出:

$$E(M)=0\times0.8+1\times0.10+2\times0.06+3\times0.03+4\times0.01=0.35 \tag{2.2}$$

也就是说,在写一篇学期论文时,电脑死机次数的期望值是0.35。当然,实际死机次数必定都是整数。在写一篇特殊学期论文时电脑死机0.35次是没有意义的!更确切地说,公式(2.2)中的计算意味着在许多这样的学期论文写作期间死机的平均次数是0.35。

取 k 个不同值的离散型随机变量 Y 的期望值表达式在重要概念2.1中给出。

贝努里随机变量的期望值。重要概念2.1中一般表达式的一个重要的特殊情形是贝努里随机变量的均值。设 G 是公式(2.1)中概率分布的贝努里随机变量。G 的期望值是:

$$E(G)=1\times p+0\times(1-p)=p \tag{2.3}$$

因此,贝努里随机变量的期望值等于 p,即它取值为"1"时的概率。

连续型随机变量的期望值。连续型随机变量的期望值也是该随机变量的可能结果的概率加权平均。由于连续型随机变量可以取连续的值,因此,连续型随机变量的期望的正式数学定义涉及微积分,它的定义在附录15.1中给出。

重要概念2.1

期望值和均值

假设随机变量 Y 取 k 个可能的值,$y_1,\cdots,y_k$,其中,y_1 表示第一个值,y_2 表示第二个值,依此类推。Y 取 y_1 的概率为 p_1,Y 取 y_2 的概率为 p,依此类推。用 $E(Y)$ 表示 Y 的期望值,它是:



$$E(Y)=y_1p_1+y_2p_2+\cdots+y_kp_k=\sum_{i=1}^{k}y_ip_i \tag{2.4}$$

其中，$\sum_{i=1}^{k}y_ip_i$ 意味着"i 取值从 1 变化到 k 时 y_ip_i 的和"。Y 的期望值也被称为 Y 的均值或 Y 的期望，通常用 μ_Y 表示。

2.2.2 方差、标准差和矩

方差和标准差测度概率分布的离散程度或"分散程度"。随机变量 Y 的方差（variance），是 Y 对其均值的离差平方的期望值，表示为 $\text{var}(Y)$，即 $\text{var}(Y)=E[(Y-\mu_Y)^2]$。

由于方差包含 Y 的平方，因此，方差的单位也就是 Y 的单位的平方，这使得方差很难解释，所以，通常用标准差（standard deviation）来测度离散程度，它是方差的平方根，表示为 σ_Y。标准差和 Y 有相同的单位。这些定义在重要概念 2.2 中进行总结。

例如，电脑死机次数 M 的方差是 M 与其均值 0.35 之差的平方的概率加权平均数。

$$\begin{aligned}\text{var}(M)&=(0-0.35)^2\times0.80+(1-0.35)^2\times0.10+(2-0.35)^2\times0.06\\&\quad+(3-0.35)^2\times0.03+(4-0.35)^2\times0.01=0.6475\end{aligned} \tag{2.5}$$

M 的标准差是方差的平方根，因此，$\sigma_M=\sqrt{0.6475}\approx0.80$。

重要概念 2.2

方差和标准差

用 σ_Y^2 表示的离散型随机变量 Y 的方差是：

$$\sigma_Y^2=\text{var}(Y)=E[(Y-\mu_Y)^2]=\sum_{i=1}^{k}(y_i-\mu_Y)^2p_i \tag{2.6}$$

Y 的标准差是 σ_Y，即方差的平方根。标准差的单位与 Y 的单位相同。

贝努里随机变量的方差。公式（2.1）中概率分布的贝努里随机变量 G 的均值是 $\mu_G=p$（公式（2.3）），所以，它的方差是：

$$\text{var}(G)=\sigma_G^2=(0-p)^2\times(1-p)+(1-p)^2\times p=p(1-p) \tag{2.7}$$

因此，贝努里随机变量的标准差为 $\sigma_G=\sqrt{p(1-p)}$。

矩。Y 的均值 $E(Y)$ 还被称为 Y 的一阶矩，Y 的平方的期望值 $E(Y^2)$ 还被称为 Y 的二阶矩。一般地说，Y^r 的期望值被称为随机变量 Y 的 r 阶矩（r^{th} moment），也就是说，Y 的 r 阶矩是 $E(Y^r)$。

就像均值是分布集中的程度测度一样，标准差是分布的离散程度的测度。$r>2$ 时的矩测度，揭示了随机变量分布形状的其他方面的信息。在本书中，分布的高阶矩（$r>2$ 时的矩）主要在重要的统计方法和经济计量方法的数学假设和推导中使用。

2.2.3 随机变量线性函数的均值和方差



这部分讨论由一个线性函数联系起来的随机变量（如 X 和 Y）。例如，考虑一个所得税方案。在这个方案下，个人的收入以 20% 的税率被征税，然后给 2 000 美元的补助金（免税的）。在这个税收方案下，将税后收入 Y 与税前收入 X 联系起来的方程为：

$$Y=2\,000+0.8X \tag{2.8}$$

于1。

表2—2　　天气状况和通勤时间的联合分布

	雨天($X=0$)	晴天($X=1$)	合计
通勤时间长($Y=0$)	0.15	0.07	0.22
通勤时间短($Y=1$)	0.15	0.63	0.78
合　计	0.30	0.70	1.00

边缘概率分布。随机变量 Y 的**边缘概率分布**(marginal probability distribution)只是其概率分布的另一个名称。这个术语用来区分单个变量 Y 的分布(边缘分布)和 Y 与另一个随机变量的联合分布。

可由 X 和 Y 的联合分布通过将 Y 取特定值的所有可能结果的概率加起来计算 Y 的边缘分布。如果 X 可取 l 个不同的值 $x_1,\cdots,x_l$,那么 Y 取特定值 y 的边缘概率是:

$$\Pr(Y=y)=\sum_{i=1}^{l}\Pr(X=x_i,Y=y) \tag{2.14}$$

例如,在表2—2中,下雨且通勤时间长的概率是15%,不下雨且通勤时间长的概率是7%,因此通勤时间长的概率(下雨或不下雨)是22%。通勤时间的边缘分布在表2—2的最后一列给出。同理,下雨的边缘概率是30%,如表2—2的最后一行所示。

2.3.2　条件分布

条件分布。以随机变量 X 取特定值为条件的随机变量 Y 的概率分布被称为**给定 X 条件下 Y 的条件分布**(conditional distribution of Y given X)。当 X 取值为 x 时,Y 取值为 y 的条件概率被记为 $\Pr(Y=y|X=x)$。

例如,如果你知道天气下雨($X=0$),那么通勤时间长($Y=0$)的概率是多少?根据表2—2可知,下雨且通勤时间短的联合概率是15%,下雨且通勤时间长的联合概率也是15%,所以,如果下雨,通勤时间长和通勤时间短的可能性是等同的。因此,以下雨($X=0$)为条件的通勤时间长($Y=0$)的概率是50%,或 $\Pr(Y=0|X=0)=0.50$。同理,下雨的边缘概率是30%,也就是说在多次通勤中有30%的机会遇到雨天。在这30%的通勤时间里,有50%的机会通勤时间长(0.15/0.30)。

一般地说,给定 $X=x$ 条件下 Y 的条件分布为:

$$\Pr(Y=y|X=x)=\frac{\Pr(X=x,Y=y)}{\Pr(X=x)} \tag{2.15}$$

例如,假如下雨,通勤时间长的条件概率为 $\Pr(Y=0|X=0)=\Pr(X=0,Y=0)/\Pr(X=0)=0.15/0.3=0.50$。

我们再看电脑死机的这个例子。假设你用图书馆的电脑打印学期论文,并且管理员在那些可以选择的电脑中为你随机地指定一台。这些电脑一半是新的,一半是旧的。由于你使用的电脑是随机指定的,因此,你所使用的电脑的新旧程度 A(如果电脑是新的,那么 $A=1$;如果电脑是旧的,那么 $A=0$)是个随机变量。假定随机变量 M 和 A 的联合分布由表2—3的A部分给出,那么,在给出了电脑的新旧程度的条件下,电脑死机次数的条件概率分布在表的B部分中给出。例如,$M=0$ 且 $A=0$ 的联合概率是0.35。由于一半电脑是旧的,假如你使用旧电脑,不死机的条件概率是 $\Pr(M=0|A=0)=\Pr(M=0,A=0)/\Pr(A=0)=0.35/0.5=0.70$,或表示为70%。相反,假设给你指定一台新电脑,不死机的条件概率是90%。根据表2—2的B部分中的条件分布,新电脑比旧电脑更不可能死机,例如,旧电脑死机3次





的概率是5%，而新电脑仅为1%。

条件期望　**给定 X 条件下 Y 的条件期望**（conditional expection of Y given X）是给定 X 条件下 Y 的条件分布的均值，也被称为**给定 X 条件下 Y 的条件均值**（conditional mean of Y given X）。也就是说，条件期望是利用给定 X 条件下 Y 的条件分布所计算的 Y 的期望值。如果 Y 取 k 个值 $y_1,\cdots,y_k$，那么给定 $X=x$ 条件下 Y 的条件均值是：

$$E(Y|X=x)=\sum_{i=1}^{k}y_iP_r(Y=y_i|X=x) \tag{2.16}$$

例如，根据表2—3中的条件分布，如果电脑是旧的，电脑死机的期望次数是 $E(M|A=0)=0\times0.70+1\times0.13+2\times0.10+3\times0.05+4\times0.02=0.56$。如果电脑是新的，电脑死机的期望次数是 $E(M|A=1)=0.14$，比旧电脑的死机次数少。

表2—3　　**电脑死机次数（M）和电脑新旧程度（A）的联合分布和条件分布**

A. 联合分布

	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	合计
旧电脑（$A=0$）	0.35	0.065	0.05	0.025	0.01	0.50
新电脑（$A=1$）	0.45	0.035	0.01	0.005	0.00	0.50
合　计	0.8	0.1	0.06	0.03	0.01	1.00

B. 已知 A 条件下 M 的条件分布

	$M=0$	$M=1$	$M=2$	$M=3$	$M=4$	合计
$\Pr(M|A=0)$	0.70	0.13	0.10	0.05	0.02	1.00
$\Pr(M|A=1)$	0.90	0.07	0.02	0.01	0.00	1.00

给定 $X=x$ 时 Y 的条件期望正好是当 $X=x$ 时 Y 的平均值。在表2—3的例子中，旧电脑死机的平均次数为0.56次，因此，假如电脑是旧的，Y 的条件期望是0.56次。同理，在新电脑中，平均死机的次数是0.14次，也就是说，假如电脑是新的，Y 的条件期望是0.14次。

累期望法则。Y 的均值就是给定 X 条件下用 X 的概率分布作为权重的 Y 的条件期望的加权平均。例如，成年人的平均身高是用男女比例作为权重的男人平均身高和女人平均身高的加权平均。用数学式表达，如果 X 取 l 个不同的值 $x_1,\cdots,x_l$，那么：

$$E(Y)=\sum_{i=1}^{l}E(Y|X=x_i)\Pr(X=x_i) \tag{2.17}$$

等式（2.17）可由公式（2.16）和公式（2.15）推导得出（见练习2.9）。

换句话说，Y 的期望就是给定 X 时 Y 的条件期望的期望，即：

$$E(Y)=E[E(Y|X)] \tag{2.18}$$

其中，公式（2.18）右边的括号内的期望，是利用给定 X 时 Y 的条件分布计算出来的，而括号外面的期望是使用 X 的边缘分布计算出来的。公式（2.18）就是著名的**累期望法则**（law of Iterated expectation）。

例如，死机次数 M 的均值，是在假定电脑是新的时 M 的条件期望与假定电脑是旧的时 M 的条件期望的加权平均数，因此，$E(M)=E(M|A=0)\times\Pr(A=0)+E(M|A=1)\times\Pr(A=1)=0.56\times0.50+0.14\times0.50=0.35$，这就是 M 边缘分布的均值，与等式（2.2）所计算的结果一样。

累期望法则的含义是，如果给定 X 时 Y 的条件均值为0，那么 Y 的均值就是0。这是公式（2.18）的直接结论：如果 $E(Y|X)=0$，那么 $=E[E(Y|X)]=E(0)=0$。换句话说，如果给

定 X 的条件下 Y 的均值为 0,那么这些条件均值的概率加权平均也一定等于 0,也就是说,Y 的均值一定为 0。

条件方差。以 X 为条件的 Y 的方差(variance of Y conditional on X)就是给定 X 条件下 Y 的条件分布的方差。用数学式表达,给定 X 条件下的 Y 的条件分布的方差为:

$$\text{var}(Y|X=x)=\sum_{i=1}^{k}[y_i-E(Y|X=x)]^2\Pr(Y=y_i|X=x) \tag{2.19}$$

例如,假如电脑是旧的,死机次数的条件方差 $\text{var}(M|A)=(0-0.56)^2\times0.70+(1-0.56)^2\times0.13+(2-0.56)^2\times0.10+(3-0.56)^2\times0.05+(4-0.56)^2\times0.02\approx0.99$,因此,假如 $A=0$,M 条件分布的标准差为$\sqrt{0.99}=0.99$。假如 $A=1$,M 的条件方差是表 2—3 的第二行中分布的方差,结果为 0.22,因此,新电脑 M 的条件标准差是$\sqrt{0.22}=0.47$。对表 2—3 中的条件分布而言,新电脑死机的期望次数(0.14)要比旧电脑的(0.56)少,而且如条件标准差所测度的一样,新电脑死机次数的分布离散程度(0.47)比旧电脑的(0.99)小。

2.3.3 独立性

如果知道一个变量的值不会提供有关另一个变量的任何信息,那么两个随机变量 X 和 Y 就是**独立分布的**(independently distribution)或者说是**独立的**(independent)。严格而言,如果给定 X 条件下 Y 的条件分布等于 Y 的边缘分布,那么 X 和 Y 就是独立的。也就是说,如果对所有 x 和 y 值,

$$\Pr(Y=y|X=x)=\Pr(Y=y)\quad(X\text{与}Y\text{独立}) \tag{2.20}$$

那么 X 和 Y 是独立分布的。

将公式(2.20)代入公式(2.15)中,我们便会得出独立随机变量以联合分布形式表示的另一种表达式。如果 X 和 Y 是独立的,那么

$$\Pr(Y=y,X=x)=\Pr(X=x)\Pr(Y=y) \tag{2.21}$$

也就是说,两个独立随机变量的联合分布是它们的边缘分布的乘积。

2.3.4 协方差与相关系数

协方差。测度两个随机变量共同变化程度的一个指标就是它们的协方差。X 和 Y 之间的**协方差**(covariance)就是期望值 $E[(X-\mu_X)(Y-\mu_Y)]$,这里 μ_X 是 X 均值,μ_Y 是 Y 均值。用 $\text{cov}(X,Y)$ 或 σ_{XY} 表示协方差,如果 X 取 l 个值,Y 取 k 个值,那么它们的协方差公式为:

$$\begin{aligned}\text{cov}(X,Y)&=\sigma_{XY}=E[(X-\mu_X)(Y-\mu_Y)]\\&=\sum_{i=1}^{k}\sum_{j=1}^{l}(x_j-\mu_X)(y_i-\mu_Y)\Pr(X=x_j,Y=y_i)\end{aligned} \tag{2.22}$$

为了解释这个公式,假设当 X 大于它的均值时(即 $X-\mu_X>0$),Y 也倾向于大于它的均值(即 $Y-\mu_Y>0$);而当 X 小于它的均值时(即 $X-\mu_X<0$),Y 也倾向于小于它的均值(即 $Y-\mu_Y<0$)。在这两种情况中,乘积项 $(Y-\mu_Y)(X-\mu_X)$ 都倾向于为正的,所以它们的协方差就是正的。相反,如果 X 和 Y 倾向于向相反方向变化(即当 Y 很小时 X 却很大,反之亦然),那么它们的协方差就是负的。最后,如果 X 和 Y 是独立的,那么它们的协方差为 0(见练习 2.9)。

相关系数。由于协方差是 X 和 Y 与其均值的离差之积,因此,它的单位就是 X 的单位乘以 Y 的单位,用起来很别扭。这个"单位"问题使得协方差的数值很难解释。

相关系数是 X 和 Y 之间相关程度的另一个测度,它解决了协方差的"单位"问题。具体





地说,X 和 Y 之间的相关系数(correlation)是 X 和 Y 的协方差除以它们各自的标准差:

$$\mathrm{corr}(X,Y)=\frac{\mathrm{cov}(X,Y)}{\sqrt{\mathrm{var}(X)\mathrm{var}(Y)}}=\frac{\sigma_{XY}}{\sigma_X\sigma_Y} \tag{2.23}$$

由于公式(2.23)中分子的单位与分母的单位相同,因此这些单位被抵消了,相关系数是无单位的。如果 $\mathrm{corr}(X,Y)=0$,那么称随机变量 X 和 Y 是不相关的(uncorrelated)。

正如在附录 2.1 中所证明的,相关系数的值总是在 -1 和 1 之间,即:

$$-1\leqslant\mathrm{corr}(X,Y)\leqslant 1 \quad (\text{相关系数不等式}) \tag{2.24}$$

相关系数与条件均值　如果 Y 的条件均值不依赖于 X,那么 Y 和 X 是不相关的,即:

$$\text{如果 } E(Y|X)=\mu_Y,\text{那么 } \mathrm{cov}(Y,X)=0 \text{ 且 } \mathrm{corr}(Y,X)=0 \tag{2.25}$$

现在我们证明这个结论。首先假设 X 和 Y 的均值都为 0,那么,$\mathrm{cov}(X,Y)=E[(Y-\mu_Y)(X-\mu_X)]=E(YX)$。根据累期望法则(公式(2.18)),由于 $E(Y|X)=0$,$E(YX)=E[E(Y|X)X]=0$,因此 $\mathrm{cov}(Y,X)=0$。方程(2.25)是通过把 $\mathrm{cov}(Y,X)=0$ 代入到公式(2.23)相关系数的定义中得到的。如果 X 和 Y 的均值不为 0,那么,先要减去它们的均值,然后应用前面的证明。

但是,如果 X 和 Y 是不相关的,那么给定 X 条件下 Y 的条件均值并不依赖于 X,这个结论就不一定成立了。换句话说,Y 的条件均值可以是 X 的函数,但 Y 和 X 却可以是不相关的,这种关系是可能的。在练习 2.10 中给出了一个这样的例子。

2.3.5　随机变量和的均值与方差

两个随机变量 X 与 Y 的和的均值等于它们的均值的和,即:

$$E(X+Y)=E(X)+E(Y)=\mu_X+\mu_Y \tag{2.26}$$

X 与 Y 的和的方差,等于它们的方差的和,加上它们协方差的 2 倍,即:

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)+2\mathrm{cov}(X,Y)=\sigma_X^2+\sigma_Y^2+2\sigma_{XY} \tag{2.27}$$

如果 X 和 Y 是独立的,那么协方差为 0,它们的和的方差等于它们的方差的和,即:

$$\mathrm{var}(X+Y)=\mathrm{var}(X)+\mathrm{var}(Y)=\sigma_X^2+\sigma_Y^2 (\text{如果 } X \text{ 和 } Y \text{ 是独立的}) \tag{2.28}$$

有关随机变量加权和的均值、方差以及协方差的重要表达式在重要概念 2.3 中给出。重要概念 2.3 中结论的推导在附录 2.1 中给出。

重要概念 2.3

随机变量和的均值、方差和协方差

设 X,Y 和 V 为随机变量,设 μ_X 和 σ_X^2 为 X 的均值和方差,σ_{XY} 为 X 和 Y 的协方差(其他变量亦如此),并设 a,b 和 c 为常数,根据均值、方差以及协方差的定义,可以得到如下公式:

$$E(a+bX+cY)=a+b\mu_X+c\mu_Y \tag{2.29}$$

$$\mathrm{var}(a+bY)=b^2\sigma_Y^2 \tag{2.30}$$

$$\mathrm{var}(aX+bY)=a^2\sigma_X^2+2ab\sigma_{XY}+b^2\sigma_Y^2 \tag{2.31}$$

$$E(Y^2)=\sigma_Y^2+\mu_Y^2 \tag{2.32}$$

$$\mathrm{cov}(a+bX+cV,Y)=b\sigma_{XY}+c\sigma_{VY} \tag{2.33}$$

$$E(YX)=\sigma_{XY}+\mu_X\mu_Y \tag{2.34}$$

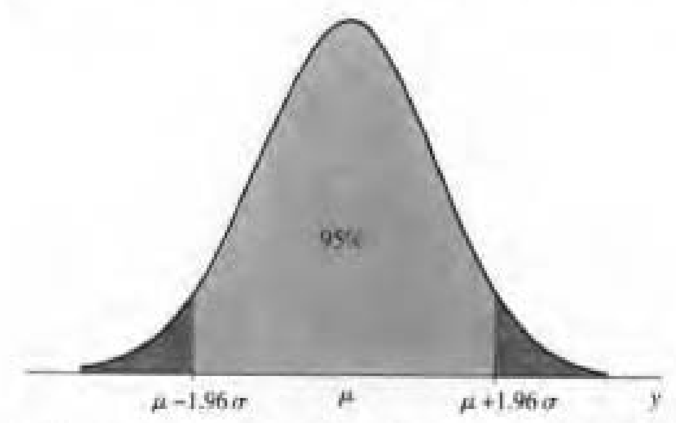
$$|\mathrm{corr}(X,Y)|\leqslant 1 \text{ 和 } |\sigma_{XY}|\leqslant\sqrt{\sigma_X^2\sigma_Y^2} (\text{相关系数不等式}) \tag{2.35}$$

2.4 正态分布、卡方分布、$F_{m,n}$分布以及学生t分布

经济计量学中最常遇到的概率分布是正态分布、卡方分布、$F_{m,n}$分布以及学生t分布。

2.4.1 正态分布

服从正态分布(normal distribution)的连续型随机变量具有图2—3中所示的类似钟形的概率密度。正态概率密度函数的具体数学定义在附录15.1中给出。如图2—3所示，均值为μ且方差为σ^2的正态密度曲线是围绕其均值对称的，且95%的概率在$\mu-1.96\sigma$和$\mu+1.96\sigma$之间。



注：均值为μ且方差为σ^2的正态概率密度函数是个钟形曲线，以μ为中心。在$\mu-1.96\sigma$和$\mu+1.96\sigma$之间的正态p.d.f.下的面积是0.95。正态分布表示为$N(\mu,\sigma^2)$。

图2—3 正态概率密度

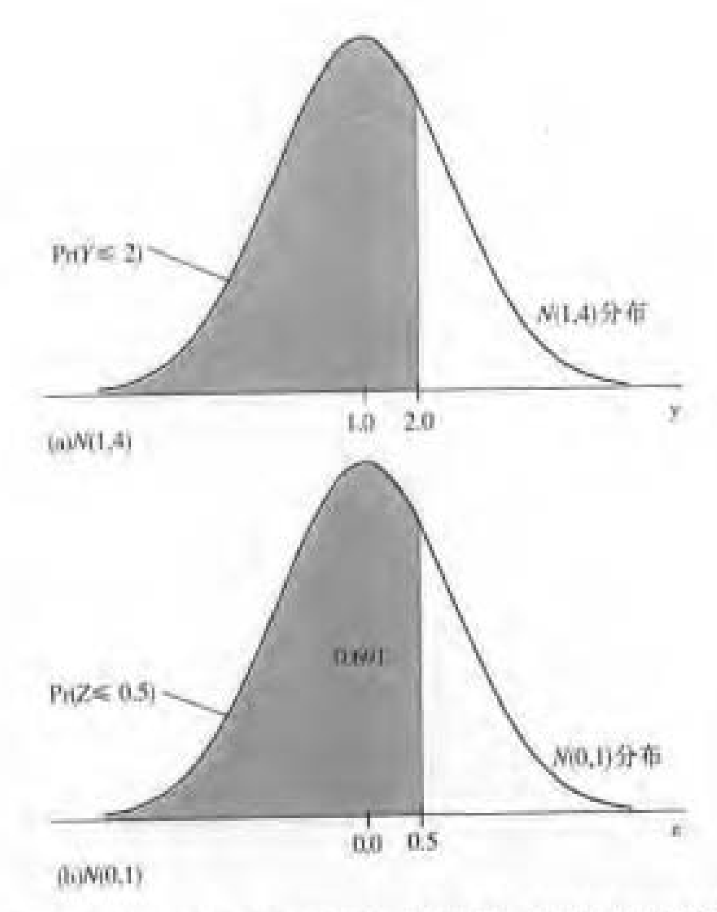
正态分布具有一些特定的记号和术语。均值为μ且方差为σ^2的正态分布简记为$N(\mu,\sigma^2)$。标准正态分布(standard normal distribution)是均值为0且方差为1的正态分布，并记为$N(0,1)$。服从$N(0,1)$的随机变量通常记为Z，标准正态累积分布函数用希腊字母Φ表示；因此，$\Pr(Z\leq c)=\Phi(c)$，这里c为常数。标准正态累积分布函数的值列在附表1中。

要计算一个具有一般均值和方差的正态变量的概率，必须首先将其标准化(standardized)，即用变量减去其均值，再除以其标准差。例如，假设Y服从$N(1,4)$分布，即Y服从均值为1且方差为4的正态分布，那么，$Y\leq2$的概率是多少呢？即图2—4(a)中阴影部分的面积是多少呢？Y的标准化形式就是Y减去它的均值后再除以它的标准差，即$(Y-1)/\sqrt{4}=\frac{1}{2}(Y-1)$。因此，随机变量$\frac{1}{2}(Y-1)$服从均值为0、方差为1的正态分布(见练习2.4)；它具有图2—4(b)所示的标准正态分布。现在，$Y\leq2$等价于$\frac{1}{2}(Y-1)\leq\frac{1}{2}(2-1)$，即，$\frac{1}{2}(Y-1)\leq\frac{1}{2}$，因此：

$$\Pr(Y\leq2)=\Pr[\frac{1}{2}(Y-1)\leq\frac{1}{2}]=\Pr(Z\leq\frac{1}{2})=\Phi(0.5)=0.691 \tag{2.36}$$

这里，0.691可以从附表1中查得。

可用同样的办法来计算正态分布的随机变量大于某个值的概率或其值落入某个范围内的概率，在重要概念2.4中总结了这些步骤。在后面的信息框"华尔街糟糕的一天"中提供



注：为了计算 $\Pr(Y\leq 2)$，标准化 Y，然后利用标准正态分布表。Y 通过减去它的均值（$\mu=1$）并除以它的标准差（$\sigma_Y=2$）进行标准化。图2—4(a)显示了 $Y\leq 2$ 的概率，图2—4(b)显示了对应的标准化 Y 之后的概率。因为标准化随机变量 $(Y-1)/2$ 是个标准正态随机变量（Z），所以，$\Pr(Y\leq 2)=\Pr((Y-1)/2\leq(2-1)/2)=\Pr(Z\leq 0.5)$。由附表1可知，$\Pr(Z\leq 0.5)=0.691$。

图2—4　计算当 Y 服从分布 $N(1,4)$ 时 $Y\leq 2$ 的概率

了累积正态分布的一个不常见的应用。

多元正态分布。正态分布可被推广来描述一组随机变量的联合分布，这种分布被称为多元正态分布(multivariate normal distribution)。如果只考虑两个变量，那么就称其为二元正态分布(bivariate normal distribution)。附录15.1中给出了二元正态分布的概率密度函数表达式，附录16.1中给出了一般的多元正态分布的概率密度函数表达式。

多元正态分布具有三个重要性质。

第一，如果 X 和 Y 服从协方差为 σ_{XY} 的二元正态分布，且 a 和 b 为常数，那么 $aX+bY$ 也服从正态分布。

$$aX+bY\sim N(a\mu_X+b\mu_Y,a^2\sigma_X^2+b^2\sigma_Y^2+2ab\sigma_{XY})\ (X,Y\text{为二元正态的}) \tag{2.37}$$

更一般地说，如果 n 个随机变量都服从多元正态分布，那么，这些变量的任何线性组合（如它们的和）也都服从正态分布。

第二，如果一组变量都服从多元正态分布，那么，每个变量的边缘分布也都是正态的（这个结论可通过令公式(2.31)中的 $a=1$ 和 $b=0$ 来得到）。

第三，如果服从多元正态分布的变量间的协方差都为0，那么，这些变量是独立的。因而，如果 X 和 Y 服从二元正态分布且 $\sigma_{XY}=0$，那么 X 和 Y 是独立的。在2.3节中阐述了如

果 X 和 Y 是独立的,那么不论它们的联合分布如何,均有 $\sigma_{XY}=0$。如果 X 和 Y 服从联合正态分布,那么反过来讲也成立。这个结论——协方差为 0 隐含着独立性——是多元正态分布的特殊性质。一般情况下,这个结论并不正确。

重要概念 2.4

正态随机变量概率的计算

假设 Y 服从均值为 μ 且方差为 σ^2 的正态分布,即 Y 服从 $N(\mu,\sigma^2)$,那么通过减去它的均值后再除以其标准差,即通过计算 $Z=(Y-\mu)/\sigma$ 可将 Y 标准化。

设 c_1 和 c_2 为两个常数且 $c_1<c_2$,并且令 $d_1=(c_1-\mu)/\sigma, d_2=(c_2-\mu)/\sigma$。则:

$$\Pr(Y\leqslant c_2)=\Pr(Z\leqslant d_2)=\Phi(d_2) \tag{2.38}$$

$$\Pr(Y\geqslant c_1)=\Pr(Z\geqslant d_1)=1-\Phi(d_1) \tag{2.39}$$

$$\Pr(c_1\leqslant Y\leqslant c_2)=\Pr(d_1\leqslant Z\leqslant d_2)=\Phi(d_2)-\Phi(d_1) \tag{2.40}$$

附表 1 中列出了正态累积分布函数 Φ。

2.4.2 卡方分布和 $F_{m,\infty}$ 分布

在统计学和经济计量学中,当我们检验某类假设时,经常会用到卡方分布和 $F_{m,\infty}$ 分布。

卡方分布(chi-squared distribution)是 m 个独立的标准正态分布随机变量的平方和的分布。这个分布依赖于 m,并称 m 为卡方分布的自由度。例如,假设 Z_1,Z_2 和 Z_3 是独立的标准正态随机变量,那么,$Z_1^2+Z_2^2+Z_3^2$ 服从自由度为 3 的卡方分布。这个分布的名称来自于用来表示它的希腊字母:自由度为 m 的卡方分布表示为 χ_m^2。

附表 3 给出了 χ_m^2 分布的部分百分位数。例如,附表 3 指出了 χ_3^2 分布的第 95 个百分位数是 7.81,因此,$\Pr(Z_1^2+Z_2^2+Z_3^2\leqslant 7.81)=0.95$。

与 χ_m^2 分布密切相关的一个分布是 $F_{m,\infty}$ 分布。**$F_{m,\infty}$ 分布**($F_{m,\infty}$ distribution)就是自由度为 m 的卡方分布的随机变量除以 m 的分布。同样,$F_{m,\infty}$ 分布就是 m 个标准正态随机变量平方和的平均数的分布。例如,如果 Z_1,Z_2 和 Z_3 是独立的标准正态随机变量,那么,$(Z_1^2+Z_2^2+Z_3^2)/3$ 服从 $F_{3,\infty}$ 分布。

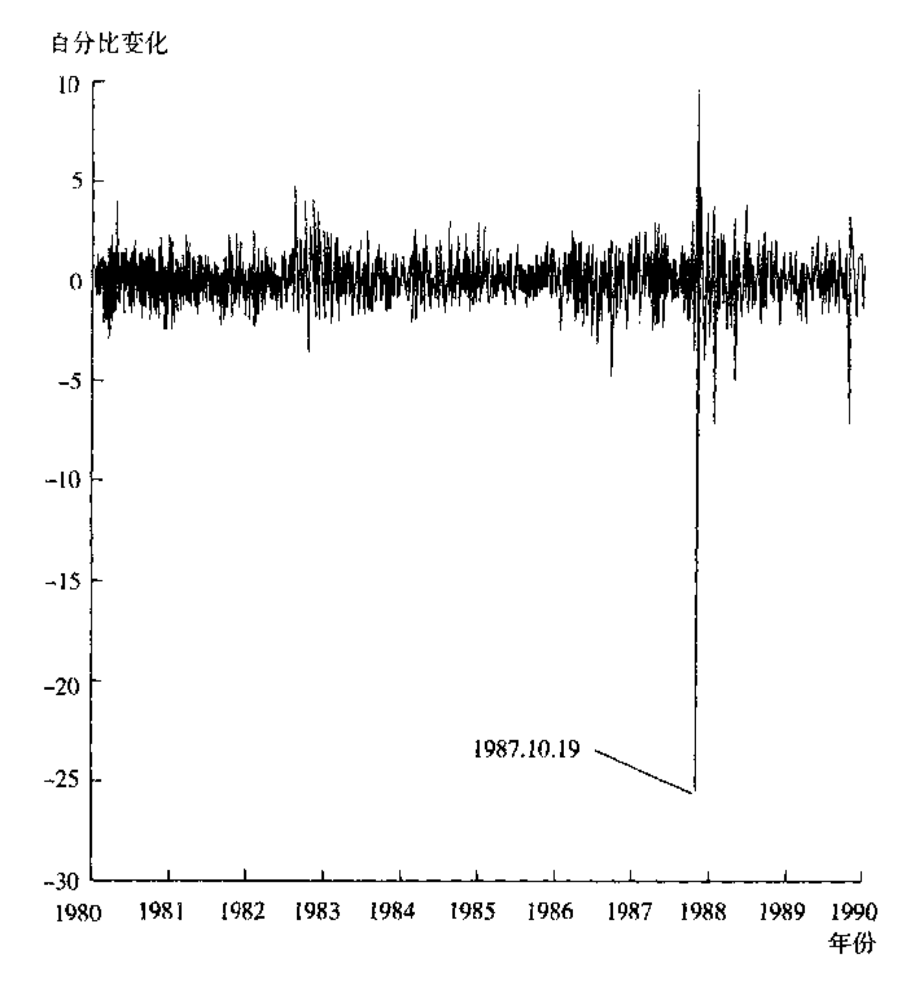
一般兴趣框

华尔街糟糕的一天

通常,在美国股票市场交易的股票的综合价值会在一日内上涨或下跌一个百分点,或者更多。这是通常情况,它根本就无法和 1987 年 10 月 9 日星期一那天所发生的一切相比。在这个"黑色星期一",道·琼斯工业平均指数(30 种大的工业股票的平均市值)下跌了 25.6%!从 1980 年 1 月 1 日到 1987 年 10 月 16 日,道氏指数日收益率的标准差(即日价格变化的百分比)是 1.16%,所以 25.6% 的下跌幅度意味着 22(25.6/1.16)个标准差的负收益。这次巨幅的下跌可从 19 世纪 80 年代道氏指数的日收益率变化图中看出,如图 2—5 所示。

如果股票收益率服从正态分布,那么至少下跌 22 个标准差的概率是 $\Pr(Z\leqslant -22)=\Phi(-22)$。在附表 1 中找不到这个值,但你可以利用计算机来计算(试一试)。这个概率等于 1.4×10^{-107},也就是 0.000…00014,小数点后共有 106 个"0"!





注:在 20 世纪 80 年代期间,"道氏"指数的日百分比平均变化是 0.05%,它的标准差是 1.16%。在 1987 年 10 月 19 日——"黑色星期一","道氏"指数下跌了 25.6%,或超过了 22 个标准差。

图 2—5 20 世纪 80 年代道·琼斯工业平均指数的日百分比变化

4×10^{-107}这个数究竟有多小呢?考虑下面的情形:

- 世界人口大约有 60 亿,因此在所有活着的人中赢得随机彩票的概率约是 60 亿分之一或 2×10^{-10}。
- 人们认为宇宙已经存在了 150 亿年或大约 5×10^{17}秒,所以从宇宙开始之时起随机地选定其中一秒的概率是 2×10^{-18}。
- 在地球表面 1 公里的大气内约有 10^{43}个分子,随机地选择一个分子的概率等于 10^{-43}。

尽管华尔街确实经历了最糟糕的一天,但它发生的概率只不过是 1.4×10^{-107}。实际上,股票收益率服从的分布要比正态分布具有更厚的尾部。换句话说,与正态分布所预示的不同,在一些日子里可能存在很大的正收益,而在另一些日子里可能存在很大的负收益。我们在第 14 章中提出了一个常被金融专家使用的股票收益率的经济计量模型,它与我们在华尔街上实际看到的非常好的或非常糟糕的日子更加一致。



附表 4 给出了 $F_{m,\infty}$ 分布的部分百分位数。例如,$F_{3,\infty}$ 分布的第 95 个百分位数是 2.60,因此,$\Pr[(Z_1^2+Z_2^2+Z_3^2)/3\leq2.60]=0.95$。$F_{3,\infty}$ 分布的第 95 个百分位数是χ_3^2 分布的第 95 个百分位数除以 3(7.81/3 = 2.60)。

(2.31)，其中，$a=b=\frac{1}{2}$，$\text{cov}(Y_1,Y_2)=0$)，$\text{var}(\overline{Y})=\frac{1}{2}\sigma_Y^2$。对于一般的 n，由于 $Y_1,\cdots,Y_n$ 是独立同分布的，对于 $i\neq j$，Y_i 与 Y_j 是独立分布的，因此 $\text{cov}(Y_i,Y_j)=0$。所以：

$$\begin{aligned}\text{var}(\overline{Y}) &= \text{var}\left(\frac{1}{n}\sum_{i=1}^{n}Y_i\right)\\ &= \frac{1}{n^2}\sum_{i=1}^{n}\text{var}(Y_i)+\frac{1}{n^2}\sum_{i=1}^{n}\sum_{j=1,j\neq i}^{n}\text{cov}(Y_i,Y_j)\\ &= \frac{\sigma_Y^2}{n}\end{aligned} \tag{2.43}$$

$\overline{Y}$ 的标准差是方差的平方根，即 $\sigma_Y/\sqrt{n}$。

综上所述，$\overline{Y}$ 的均值、方差和标准差分别为：

$$E(\overline{Y})=\mu_Y \tag{2.44}$$

$$\text{var}(\overline{Y})=\sigma_{\overline{Y}}^2=\frac{\sigma_Y^2}{n} \tag{2.45}$$

$$\text{std. dev}(\overline{Y})=\sigma_{\overline{Y}}=\frac{\sigma_Y}{\sqrt{n}} \tag{2.46}$$

不论 Y_i 的分布形式如何，上面的结论都成立。也就是说，要使公式(2.44)、公式(2.45)和公式(2.46)成立，并不要求 Y_i 的分布取某种特定的形式，比如说正态分布。

符号 $\sigma_{\overline{Y}}^2$ 表示样本均值 $\overline{Y}$ 的抽样分布的方差。相反，σ_Y^2 是每个个体 Y_i 的方差，即所抽取观测值的总体分布的方差。同样，符号 $\sigma_{\overline{Y}}$ 表示 $\overline{Y}$ 抽样分布的标准差。

当 Y 服从正态分布时，$\overline{Y}$ 的抽样分布。假定 $Y_1,\cdots,Y_n$ 是采用独立同分布抽样方法从总体 $N(\mu_Y,\sigma_Y^2)$ 中抽取的。根据公式(2.37)，n 个服从正态分布的随机变量的和也服从正态分布。由于 $\overline{Y}$ 的均值是 μ_Y，且 $\overline{Y}$ 的方差为 σ_Y^2/n，这意味着，如果 $Y_1,\cdots,Y_n$ 是采用独立同分布抽样方法从 $N(\mu_Y,\sigma_Y^2)$ 中抽取的，那么 $\overline{Y}$ 便服从分布 $N(\mu_Y,\sigma_Y^2/n)$。

2.6 抽样分布的大样本逼近

抽样分布在统计学和经济计量学方法的发展中发挥了重要作用，因此在数学意义上理解 $\overline{Y}$ 的抽样分布的含义是非常重要的。刻画抽样分布有两种方法：一种是“精确”方法，另一种是“逼近”方法。

“精确”方法要求推导出对任意的 n 值都成立的抽样分布表达式。对任意的 n，精确描述 $\overline{Y}$ 分布的抽样分布被称为 $\overline{Y}$ 的**精确分布**(exact distribution)或**有限样本分布**(finite—sample distribution)。例如，如果 Y 服从正态分布，而且 $Y_1,\cdots,Y_n$ 是独立同分布的，那么(如同 2.5 节中所讨论的)，$\overline{Y}$ 的精确分布是均值为 μ_Y 且方差为 σ_Y^2/n 的正态分布。但是，如果 Y 不服从正态分布，那么一般而言 $\overline{Y}$ 的精确抽样分布是非常复杂的，并且取决于 Y 的分布。

这里的“逼近”方法，应用了依赖于大样本容量的抽样分布的逼近理论。抽样分布的大样本逼近通常被称为**渐近分布**(asymptotic distribution)。之所以称之为“渐近的”，是因为这种逼近在 $n\to\infty$ 的极限处是精确的。正如我们在本节中所看到的，这些逼近会是非常准确的，即使样本容量只是 $n=30$ 个观测值。由于实际中经济计量学所使用的样本容量通常是数以百计或千计的，因此，能够依靠这些渐近分布为精确抽样分布提供非常好的逼近。

本节介绍当样本容量很大时用来进行逼近抽样分布的两个重要工具：大数定律和中心







极限定理。大数定律表明，当样本容量很大时，$\bar{Y}$将会以非常高的概率接近于μ_Y。中心极限定理表明，当样本容量很大时，标准化的样本均值$(\bar{Y}-\mu_Y)/\sigma_{\bar{Y}}$的抽样分布是渐近正态分布。

虽然精确的抽样分布是复杂的，而且依赖于Y的分布，但是渐近分布是简单的。此外，值得注意的是，$(\bar{Y}-\mu_Y)/\sigma_{\bar{Y}}$的渐近正态分布并不依赖于$Y$的分布。这个正态渐近分布提供了极大的简化，并成为全书所使用的回归理论的基础。

2.6.1 大数定律与一致性

大数定律(law of large numbers)指出，在一般条件下，当n很大时，$\bar{Y}$将会以非常高的概率接近于μ_Y。大数定律有时被称为"均值定律"。当把大量的均值相同的随机变量放在一起取平均数时，大的数值平衡了小的数值，而且它们的样本均值接近于它们的共同均值。

例如，考虑我们的学生通勤实验的一种简单情况，实验里她只记录下她的通勤时间是短的(少于20分钟)还是长的。如果她在第i个随机选择的日子中的通勤时间短，那么令$Y_i=1$；如果通勤时间长，那么$Y_i=0$。因为她使用简单随机抽样，所以$Y_1,\cdots,Y_n$是独立同分布的。因此，$Y_i(i=1,\cdots,n)$是贝努里随机变量的独立同分布抽样，这里(见表2—2)$Y_i=1$的概率是0.78。由于贝努里随机变量的期望是它的成功概率，因此，$E(Y_i)=\mu_Y=0.78$。样本均值$\bar{Y}$就是她的样本中通勤时间短的日子的比例。

图2—6显示了各种不同样本容量n条件下的$\bar{Y}$的抽样分布。当$n=2$时(见图2—6(a))，$\bar{Y}$只能取三个值，即0，$\frac{1}{2}$和1(没有一次通勤时间短、一次通勤时间短和两次通勤时间都短)，其中没有一个特别接近总体的真实比例0.78。但是，随着n的增大(见图2—6(b)—图2—6(d))，$\bar{Y}$取更多的值，抽样分布紧密地集中在μ_Y附近。

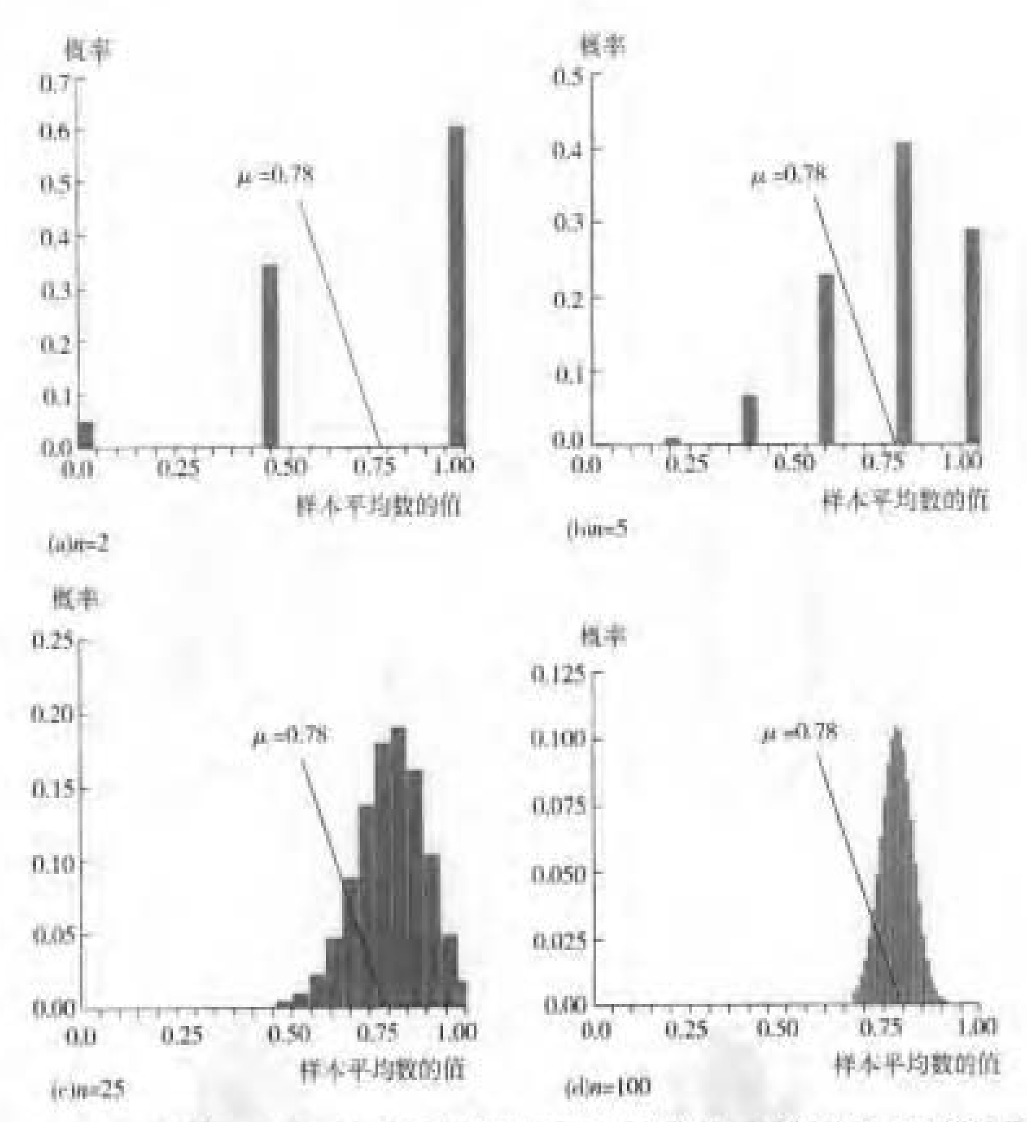
随着n的增大，$\bar{Y}$以不断增大的概率接近于μ_Y的性质被称为**依概率收敛**(convergence in probability)，或更精确地被称为**一致性**(consistency)(见重要概念2.6)。大数定律指出，在一定条件下，$\bar{Y}$依概率收敛于μ_Y，或者说，$\bar{Y}$与μ_Y是一致的。

大数定律的条件(这些条件本书中会经常用到)为：$Y_i(i=1,\cdots,n)$是独立同分布的，且Y_i的方差σ_Y^2是有限的。在15.2节中我们详细阐述了这些条件的数学意义，并证明了大数定律。如果数据是通过简单随机抽样方法搜集上来的，那么独立同分布假设就会成立。方差有限假设说明，极端大的Y_i的值很少被观测到，否则样本均值就会是不可靠的。本书采用的这个假设是合理的。例如，学生的通勤时间就存在一个上限(如果交通瘫痪，她可能会停车步行)，因此通勤时间分布的方差是有限的。

2.6.2 中心极限定理

中心极限定理(central limit theorem)指出，在一般条件下，当n较大时，$\bar{Y}$的分布可充分地逼近于正态分布。前面我们讲过，$\bar{Y}$的均值是μ_Y，且它的方差是$\sigma_{\bar{Y}}^2=\sigma_Y^2/n$。根据中心极限定理，当$n$很大时，$\bar{Y}$的分布近似服从$N(\mu_Y,\sigma_{\bar{Y}}^2)$。如2.5节结尾部分所讨论的，当样本从正态分布总体$N(\mu_Y,\sigma_Y^2)$中被随机地抽取时，$N(\mu_Y,\sigma_{\bar{Y}}^2)$就是$\bar{Y}$的精确的分布形式。中心极限定理表明，当$n$值很大时，即使$Y_1,\cdots,Y_n$本身并不服从正态分布，那么，这个结论也同样是近似正确的。

由图2—6可大致地看出$\bar{Y}$的分布收敛于钟形的正态近似。不过，对于大的n值，由于分布变得非常紧凑，需要睁大眼睛才能看得清。如果使用放大镜或利用一些别的方法放大



注：这些分布是当$p=\Pr(Y_i=1)=0.78$（快速通勤的概率为78%）时，n个独立的贝努里随机变量的样本均值$\overline{Y}$的抽样分布。$\overline{Y}$的抽样分布方差随n变大而减小，因此，随着样本容量n的增加，抽样分布变得更紧密地集中在它的均值$\mu=0.78$附近。

图2—6　n个贝努里随机变量的样本均值的抽样分布

或扩大图中的横坐标轴，就会更容易地看出$\overline{Y}$的分布形状。

重要概念2.6

依概率收敛、一致性和大数定律

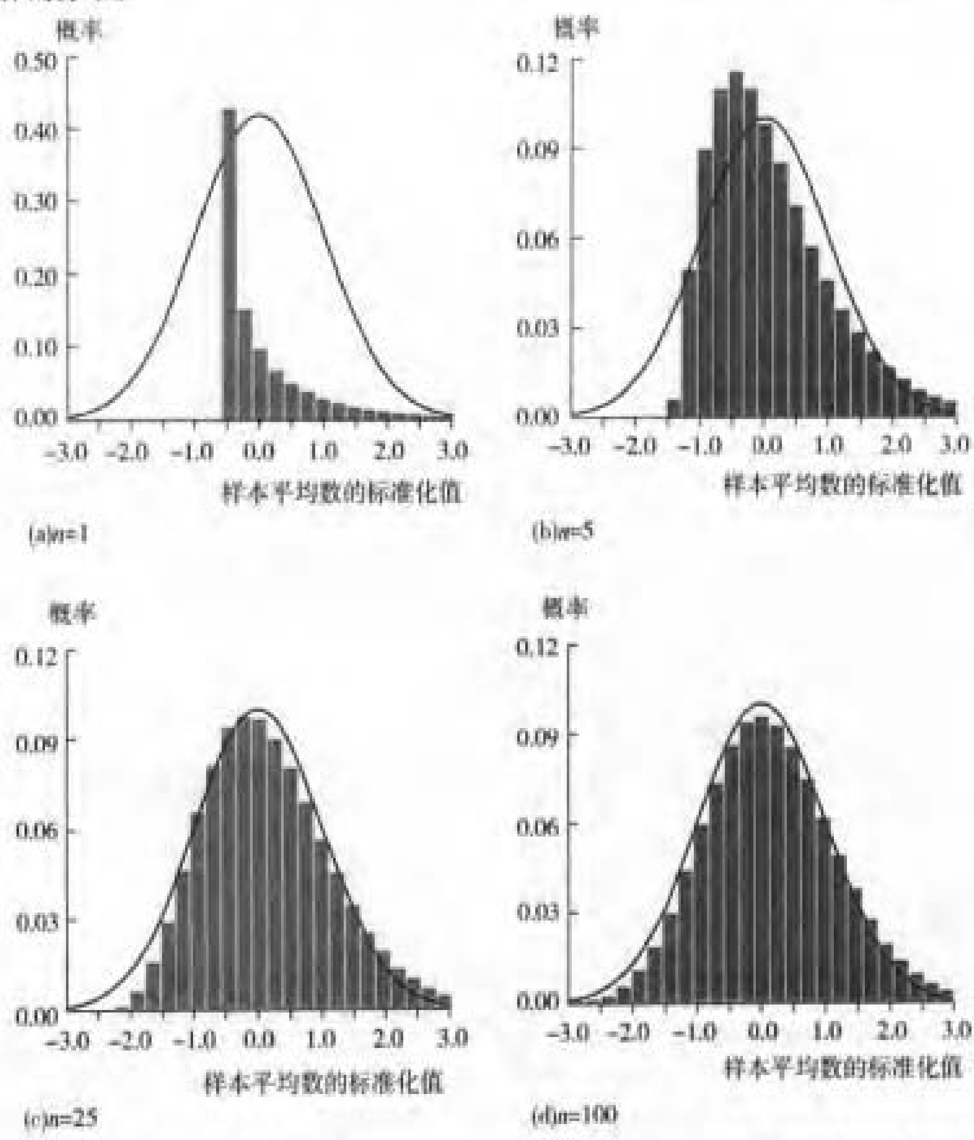
对于任意的常数$c>0$，随n增大时，如果$\overline{Y}$位于(μ_Y-c,μ_Y+c)区域内的概率任意地接近1，那么，我们就说样本均值$\overline{Y}$依概率收敛于μ_Y（或者说，$\overline{Y}$与μ_Y是一致的），记为$\overline{Y}\xrightarrow{p}\mu_Y$。

大数定律指出：如果$Y_i(i=1,\cdots,n)$是独立同分布的，且$E(Y_i)=\mu_Y$，$\mathrm{var}(Y_i)=\sigma_Y^2<\infty$，那么$\overline{Y}\xrightarrow{p}\mu_Y$。

使分布形态更为清晰的一种规范方法就是标准化，即减去$\overline{Y}$的均值再除以它的标准

可能要求 $n=30$ 或更大。

对总体分布而言，这一观点在图 2—8 中得到了说明（见图 2—8(a)），它明显不同于贝努里分布。这个分布有一个长的右尾（它是"右偏"的）。中心化和调整坐标刻度之后，$n=5$、$n=25$ 和 $n=100$ 的 $\bar{Y}$ 的抽样分布分别如图 2—8(b)、图 2—8(c) 和图 2—8(d) 所示。虽然 $n=25$ 的抽样分布接近于钟形，但正态近似仍然具有不可忽视的不完美性。不过，当 $n=100$ 时，正态近似是相当好的。事实上，当 $n\geqslant 100$ 时，$\bar{Y}$ 分布的正态近似非常适合于各种广泛总体的分布。



注：图 2—8(a) 显示了来自于一个有偏的（不对称的）总体分布中 n 个抽样的标准化样本均值的抽样分布。当 n 较小时（$n=5$），抽样分布像总体分布一样是有偏的。但当 n 很大时（$n=100$），如中心极限定理所预测的那样，抽样分布被一个标准正态分布（实线）很好地近似。

图 2—8　取自于一个有偏分布的 n 个抽样的标准化样本均值的分布

中心极限定理是个重要的结论。当 n 很小时，图 2—7、图 2—8(b) 和图 2—8(c) 中 $\bar{Y}$ 的分布是很复杂的，且彼此差异很大；而当 n 较大时，图 2—7(d) 和图 2—8(d) 中的分布却很简单，并且有着惊人的相似。由于当 n 增大时，$\bar{Y}$ 的分布接近于正态分布，因此 $\bar{Y}$ 的分布被称为是渐近正态分布的（asymptotic normally distributed）。



中心极限定理的作用和正态近似的方便性，使得它们的理论很容易与现实的广泛应用相结合，从而使它们成为了现代应用统计学的重要基础。中心极限定理要点的总结在重要概念 2.7 中给出。

重要概念 2.7

中心极限定理

假设 $Y_1,\cdots,Y_n$ 是独立同分布的，且 $E(Y_i)=\mu_Y$，$\text{var}(Y_i)=\sigma_Y^2$，其中，$0<\sigma_Y^2<\infty$。当 $n\to\infty$ 时，$(\bar{Y}-\mu_Y)/\sigma_{\bar{Y}}$（其中，$\sigma_{\bar{Y}}^2=\sigma_Y^2/n$）的分布会被标准正态分布很好地近似替代。

总结

1. 随机变量取不同值时的概率，由累积分布函数、概率分布函数（对于离散型随机变量而言）和概率密度函数（对于连续型随机变量而言）来描述。

2. 随机变量 Y 的期望值（也称为它的均值 μ_Y）是它的概率加权平均值，表示为 $E(Y)$。Y 的方差是 $\sigma_Y^2=E[(Y-\mu_Y)^2]$，$Y$ 的标准差是其方差的平方根。

3. 两个随机变量 X 和 Y 的联合概率由它们的联合概率分布来描述。给定 $X=x$ 条件下 Y 的条件概率分布，就是以 X 取值 x 为条件的 Y 的概率分布。

4. 正态分布的随机变量具有图 2—3 中的钟形的概率密度曲线。要计算与正态随机变量相联系的概率，首先要将变量标准化，然后使用附表 1 中列出的标准正态累积分布表。

5. 简单随机抽样生成 n 个独立同分布（i. i. d.）的随机观测值 $Y_1,\cdots,Y_n$。

6. 样本均值 $\bar{Y}$ 随着被随机选择的样本的变化而变化，因而它是个具有抽样分布特征的随机变量。如果 $Y_1,\cdots,Y_n$ 是独立同分布的，那么：

a. $\bar{Y}$ 的抽样分布均值为 μ_Y，方差为 $\sigma_{\bar{Y}}^2=\sigma_Y^2/n$；

b. 大数定律表明，$\bar{Y}$ 依概率收敛于 μ_Y；

c. 中心极限定理表明，$\bar{Y}$ 的标准化形式 $(\bar{Y}-\mu_Y)/\sigma_{\bar{Y}}$，在 n 值很大时，服从标准正态分布（$N(0,1)$ 分布）。

重要术语

结果　概率　样本空间　事件　离散型随机变量　连续型随机变量　概率分布　累积概率分布　累积分布函数（c. d. f.）　贝努里随机变量　贝努里分布　概率密度函数（p. d. f.）　密度函数　密度　期望值　均值　方差　标准差　分布的矩　联合概率分布　边缘概率分布　条件分布　条件期望　条件均值　累期望法则　条件方差　独立性　协方差和相关性　不相关的　正态分布　标准正态分布　标准化变量　多元正态分布　二元正态分布　卡方分布　$F_{m,\infty}$ 分布　学生 t 分布　简单随机抽样　标准化随机变量　总体　同分布的　独立同分布（i. i. d.）　抽样分布　精确分布　渐近分布　大数定律　依概率收敛　一致性　中心极限定理　渐近正态分布



复习概念

2.1　本章中用到的随机变量的例子包括：(a) 你遇到的下一个陌生人的性别；(b) 电脑

死机次数；(c)通勤到学校花费的时间；(d)你在图书馆被指定使用的电脑是新的还是旧的；(e)是否下雨。请解释为什么每一个变量都可以被看做是随机的。

2.2　假定随机变量 X 和 Y 是独立的，并且知道它们的分布。请解释为什么知道 X 的值但不会因而知道任何关于 Y 值的信息。

2.3　设 X 表示你的家乡在某个给定月份中的降雨量，Y 表示同一月份洛杉矶出生的婴儿数。请问：X 和 Y 是相互独立的吗？请给予说明。

2.4　一个经济计量学课程的班级里有 80 名学生，这些学生的平均体重是 145 磅，从这个班级中随机选择 4 名学生作为样本，并计算出他们的平均体重。在这个样本中，学生的平均体重会等于 145 磅吗？如果会，为什么？如果不会，为什么？结合这个例子解释为什么样本均值 $\bar{Y}$ 是个随机变量。

2.5　假设 $Y_1,\cdots,Y_n$ 是独立同分布的随机变量，服从 $N(1,4)$ 分布。绘制当 $n=2$ 时 $\bar{Y}$ 的概率密度草图。当 $n=10$ 和 $n=100$ 时重复这一过程。用语言描述这些概率密度的差别。你的答案和大数定律之间有什么关系吗？

2.6　假设 $Y_1,\cdots,Y_n$ 是独立同分布的随机变量，服从图 2—8(a) 中给出的概率分布，现在你想要计算 $\Pr(\bar{Y}\leq 0.1)$。如果 $n=5$，使用正态近似合理吗？如果 $n=25$ 或 $n=100$ 呢？请给予解释和说明。

练习

带“＊”号的练习题的答案可以在网址 www.aw.com/stock_watson 上找到。

*2.1　使用表 2—2 中给出的概率分布计算：(a) $E(Y)$ 和 $E(X)$；(b) σ_X^2 和 σ_Y^2；(c) σ_{XY} 和 $\operatorname{corr}(X,Y)$。

2.2　使用表 2—2 中的随机变量 X 和 Y，考虑另外两个新的随机变量 $W=3+6X$ 和 $V=20-7Y$。请计算：(a) $E(W)$ 和 $E(V)$；(b) σ_W^2 和 σ_V^2；(c) σ_{WV} 和 $\operatorname{corr}(W,V)$。

2.3　表 2—4 给出了根据 1990 年美国人口普查得出的就业或失业（正在找工作）的劳动适龄人口中，大学毕业生和就业状况之间的联合概率分布数据。

表 2—4　　1990 年 25～64 岁的美国人口中大学毕业生和就业状况的联合分布

	失业($Y=0$)	就业($Y=1$)	总计
非大学毕业生($X=0$)	0.045	0.709	0.754
大学毕业生($X=1$)	0.005	0.241	0.246
总　计	0.050	0.950	1.000

*a. 计算 $E(Y)$。

b. 失业率是劳动力中失业人口所占的比例。证明失业率等于 $1-E(Y)$。

*c. 计算 $E(Y|X=1)$ 和 $E(Y|X=0)$。

d. 分别计算大学毕业生和非大学毕业生的失业率。

*e. 从这个总体中随机选出一个人，而且这个人处于失业状态。请问：这个劳动者为大学毕业生的概率是多少？为非大学毕业生的概率又是多少？

f. 受教育程度和就业状况之间是独立的吗？请解释理由。

2.4　随机变量 Y 的均值等于 1，方差等于 4，设 $Z=(Y-1)/2$。证明：$\mu_Z=0,\sigma_Z^2=1$。

2.5　计算下列概率：



$$=b^2E[(Y-\mu_Y)^2]$$
$$=b^2\sigma_Y^2。$$

为了推导公式(2.31),我们仍然利用方差的定义:

$$\begin{aligned}\mathrm{var}(aX+bY)&=E\{[aX+bY-(a\mu_X+b\mu_Y)]^2\}\\&=E\{[a(X-\mu_X)+b(Y-\mu_Y)]^2\}\\&=E[a^2(X-\mu_X)^2]+2E[ab(X-\mu_X)(Y-\mu_Y)]+E[b^2(Y-\mu_Y)^2]\\&=a^2\mathrm{var}(X)+2ab\mathrm{cov}(X,Y)+b^2\mathrm{var}(Y)\\&=a^2\sigma_X^2+2ab\sigma_{XY}+b^2\sigma_Y^2\end{aligned}\tag{2.47}$$

其中,第二个公式是通过合并同类项得到的;第三个公式是通过展开二次项得到的;第四个公式是根据方差及协方差的定义得到的。

现在推导公式(2.32)。由于$E(Y-\mu_Y)=0$,因此:

$$E(Y^2)=E\{[(Y-\mu_Y)+\mu_Y]^2\}=E[(Y-\mu_Y)^2]+2\mu_YE(Y-\mu_Y)+\mu_Y^2=\sigma_Y^2+\mu_Y^2$$

为了推导公式(2.33),我们利用协方差的定义:

$$\begin{aligned}\mathrm{cov}(a+bX+cV,Y)&=E\{[a+bX+cV-E(a+bX+cV)][Y-\mu_Y]\}\\&=E\{[b(X-\mu_X)+c(V-\mu_V)][Y-\mu_Y]\}\\&=E\{[b(X-\mu_X)][Y-\mu_Y]\}+E\{[c(V-\mu_V)][Y-\mu_Y]\}\\&=b\sigma_{XY}+c\sigma_{VY}\end{aligned}\tag{2.48}$$

$b\sigma XY+c\sigma VY$就是公式(2.33)。

现在我们推导公式(2.34)。

$$\begin{aligned}E(XY)&=E\{[(X-\mu_X)+\mu_X][(Y-\mu_Y)+\mu_Y]\}\\&=E[(X-\mu_X)(Y-\mu_Y)]+\mu_XE(Y-\mu_Y)+\mu_YE(X-\mu_X)+\mu_X\mu_Y\\&=\sigma_{XY}+\mu_X\mu_Y\end{aligned}$$

现在,我们来证明公式(2.35)中相关系数的不等式,即$|\mathrm{corr}(X,Y)|\leqslant1$。设$a=-\sigma_{XY}/\sigma_X^2$,$b=1$,应用公式(2.31),则:

$$\begin{aligned}\mathrm{var}(aX+Y)&=a^2\sigma_X^2+\sigma_Y^2+2a\sigma_{XY}\\&=(-\sigma_{XY}/\sigma_X^2)^2\sigma_X^2+\sigma_Y^2-2(-\sigma_{XY}/\sigma_X^2)\sigma_{XY}\\&=\sigma_Y^2-\sigma_{XY}^2/\sigma_X^2\end{aligned}\tag{2.49}$$

因为$\mathrm{var}(aX+Y)$是方差,它不可能是负的,所以由公式(2.49)中的最后一行可知,$\sigma_Y^2-\sigma_{XY}^2/\sigma_X^2\geqslant0$。重新整理这个不等式,得到:

$$\sigma_{XY}^2\leqslant\sigma_X^2\sigma_Y^2\quad(\text{协方差不等式})\tag{2.50}$$

这个协方差不等式意味着$\sigma_{XY}^2/(\sigma_X^2\sigma_Y^2)\leqslant1$,或者说,$|\sigma_{XY}/(\sigma_X\sigma_Y)|\leqslant1$,这(利用相关系数的定义)就证明了相关系数不等式$|\mathrm{corr}(X,Y)|\leqslant1$。

第3章 统计学知识复习

统计学是利用数据来了解我们身边世界的一门学科。统计学的工具可以帮助我们知晓我们感兴趣的总体分布的未知特征。例如，刚毕业的大学生的收入分布的均值是多少？男性职工和女性职工的平均收入是否有差别？如果有，差别是多大？

这些问题与劳动者总体收入的分布有关。回答这些问题的一个方法是对劳动者总体进行一次彻底的调查，记录每个劳动者的收入，进而得到收入的总体分布。但实际上，这样的全面调查是极其昂贵的。美国惟一的一个全面调查是10年一度的人口普查。仅2000年的美国人口普查就耗资100亿美元，设计普查表、进行管理和实施调查、收集与分析数据的过程就花费了10年时间。尽管付出了如此大的努力，还是有许多人口被遗漏，没有被调查。因此我们需要一种不同的、更实用的方法。

统计学的一个重要观点是人们能够通过对从总体中选择随机样本的认识来了解总体的分布情况。比如，不用调查美国人口总体，我们只需通过简单随机抽样的方法从总体人口中随机选择一部分个体，如1 000人，然后利用统计方法通过对这个样本进行分析，我们就可以得到关于整个总体特征的试验性结论，即进行统计推断。

有三种统计方法在整个经济计量学中被普遍使用，即估计方法、假设检验方法和置信区间方法。所谓估计方法，就是从样本数据中计算出总体分布未知特征的"最佳猜测值"，如样本的均值。假设检验方法，需要提出一个关于总体特征的特定的假设，然后利用样本证据判定它是否是真的。置信区间方法，是用一组样本数据估计未知总体特征的某一区间或范围。在第3.1节、第3.2节和第3.3节中复习了关于未知总体均值统计推断的参数估计、假设检验和置信区间。

经济学中许多令人感兴趣的问题，都涉及两个或两个以上变量之间的关系或不同总体之间的比较。例如，刚毕业的大学生中男女平均收入之间是否存在差异？第3.1节至第3.3节中介绍的单个总体均值估计的方法，被推广应用到比较两个不同总体的均值，这部分内容在第3.4节介绍。第3.5节则应用这些方法，研究了男女大学毕业生收入之间的"性别差异"的证据。第3.6节以样本相关关系和散点图的讨论作为本章的结束。

3.1 总体均值的估计

假如你想知道一个总体 Y 的均值 μ_Y，如刚毕业的女大学生的平均收入，那么，估计这个均值的一个很自然的方法，就是从 n 个独立同分布（i. i. d. ）的样本观测值 $Y_1,\cdots,Y_n$ 中计算样本的平均值 $\overline{Y}$（以前我们讲过，如果 $Y_1,\cdots,Y_n$ 是通过简单随机抽样方法获得的，那么它们就是独立同分布的）。这一节我们讨论 μ_Y 的估计和作为 μ_Y 估计量的 $\overline{Y}$ 的一些性质。

3.1.1 估计量及其性质

估计量　用样本平均值 $\overline{Y}$ 估计 μ_Y 是一种很自然的方法，但它不是惟一的方法。例如，另一种估计 μ_Y 的方法是简单地用第一个观测值 Y_1 来估计。$\overline{Y}$ 和 Y_1 都是设计用来估计 μ_Y 的数据的函数，根据重要概念 3.1 中术语的解释，二者都是 μ_Y 的估计量。当我们用重复抽样的方法估计其值时，$\overline{Y}$ 和 Y_1 的值会因样本不同而不同（它们产生不同的估计值）。因此，估计量 $\overline{Y}$ 和 Y_1 都具有抽样分布。实际上，μ_Y 有很多个估计量，$\overline{Y}$ 和 Y_1 只是其中的两个。

重要概念 3.1

估计量和估计值

估计量（estimator）是从总体中随机抽取的样本数据的函数。**估计值**（estimate）是利用特定样本数据实际计算出的估计量的数值。由于样本选择具有随机性，因此估计量也是随机变量，而估计值是非随机的数值。

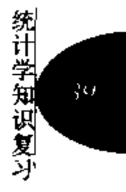
由于有很多个可能的估计量，因此，怎样判定一个估计量“好”于另一个估计量呢？因为估计量是随机变量，所以，这个问题也可以更为精确地表述为：估计量的抽样分布的理想特征是什么？一般地说，我们希望估计量至少在平均意义上尽可能地逼近未知的真值，换句话说，我们希望估计量的抽样分布尽可能紧密地集中在未知值的周围。基于这种理解，我们引出了估计量的三个特定的理想性质：无偏性（缺少偏差）、一致性和有效性。

无偏性　假设你通过重复随机抽样多次计算估计量的值，那么，在平均的意义上你希望能够得到你所要求得到的答案，这种希望是合理的。因此，估计量的一个理想性质就是它的抽样分布的均值等于 μ_Y，如果是这样的话，估计量就是无偏的。

用数学语言来表述，设 $\hat{\mu}_Y$ 表示 μ_Y 的某个估计量，如 $\overline{Y}$ 或 Y_1。如果 $E(\hat{\mu}_Y)=\mu_Y$，那么，估计量 $\hat{\mu}_Y$ 是无偏的，这里 $E(\hat{\mu}_Y)$ 是 $\hat{\mu}_Y$ 的抽样分布均值；否则，$\hat{\mu}_Y$ 是有偏的。

一致性　估计量 $\hat{\mu}_Y$ 的另一个理想性质是，当样本容量很大时，由样本的随机变化所引起的 μ_Y 值的不确定性很小。更确切地讲，$\hat{\mu}_Y$ 的理想性质是，当样本容量增大时，它在真值 μ_Y 附近很小区间内的概率接近于 1，即 $\hat{\mu}_Y$ 对 μ_Y 是一致的（见重要概念 2.6）。

方差与有效性　假如你有两个备选的估计量 $\hat{\mu}_Y$ 和 $\tilde{\mu}_Y$，而且它们都是无偏的，那么，你如何在它们之间做出选择呢？一种方法就是选择抽样分布最密集的估计量，也就是说，在 $\hat{\mu}_Y$ 和 $\tilde{\mu}_Y$ 之间选择方差最小的那个估计量。如果 $\hat{\mu}_Y$ 的方差比 $\tilde{\mu}_Y$ 的方差小，那么称 $\hat{\mu}_Y$ 比 $\tilde{\mu}_Y$ 更有效。“有效性”这个术语来自于这样的理解：如果 $\hat{\mu}_Y$ 的方差比 $\tilde{\mu}_Y$ 的方差小，那么说明 $\hat{\mu}_Y$ 比 $\tilde{\mu}_Y$ 更有效地利用了数据中的信息。



偏差、一致性和有效性的总结在重要概念 3.2 中。



3.1.2 $\overline{Y}$的性质

当用偏差、一致性和有效性这三个准则进行判断时，$\overline{Y}$作为μ_Y的估计量表现如何呢？

偏差和一致性　关于$\overline{Y}$的抽样分布，我们已经在第2.5节和第2.6节中做了分析。在第2.5节中已经证明，$E(\overline{Y})=\mu_Y$，因此$\overline{Y}$是μ_Y的无偏估计量。同理，大数定律（见重要概念2.6）结果表明：$\overline{Y}\xrightarrow{p}\mu_Y$，因此$\overline{Y}$也是一致性估计量。

有效性　$\overline{Y}$的有效性怎么样呢？由于有效性要求对估计量进行比较，因此，我们需要设定与$\overline{Y}$相比较的一个或几个估计量。

我们先从比较估计量Y_1和$\overline{Y}$的有效性开始。由于$Y_1,\cdots,Y_n$是独立同分布的，Y_1的抽样分布的均值是$E(Y_1)=\mu_Y$，因此，Y_1是μ_Y的一个无偏估计量。它的方差是$\text{var}(Y_1)=\sigma_Y^2$。由2.5节可知，$\overline{Y}$的方差等于$\sigma_Y^2/n$。因此，对于$n\geqslant2$，$\overline{Y}$的方差要比$Y_1$的方差小，也就是说，$\overline{Y}$是个比$Y_1$更有效的估计量。所以，根据有效性准则，应该用$\overline{Y}$代替$Y_1$。显然，估计量$Y_1$给你留下的印象是个较差估计量——为什么你要特意收集n个观测值的样本却只保留第一个观测值，而把所有其他的观测值舍弃掉呢？有效性概念提供了一种正式方法证明$\overline{Y}$比Y_1是个更理想的估计量。

一个不太明显的较差的估计量的情况会怎么样呢？考虑用$\frac{1}{2}$和$\frac{3}{2}$交替作为权重对观测值进行加权平均：

$$\widetilde{Y}=\frac{1}{n}\left(\frac{1}{2}Y_1+\frac{3}{2}Y_2+\frac{1}{2}Y_3+\frac{3}{2}Y_4+\cdots+\frac{1}{2}Y_{n-1}+\frac{3}{2}Y_n\right)\tag{3.1}$$

重要概念3.2

偏差、一致性和有效性

设$\hat{\mu}_Y$为μ_Y的估计量，那么：

■ $\hat{\mu}_Y$的偏差(bias)为$E(\hat{\mu}_Y)-\mu_Y$。

■ 如果$E(\hat{\mu}_Y)=\mu_Y$，那么$\hat{\mu}_Y$是μ_Y的**无偏估计量**(unbiased estimator)。

■ 如果$\hat{\mu}_Y\xrightarrow{p}\mu_Y$，那么$\hat{\mu}_Y$是$\mu_Y$的**一致估计量**(consistent estimator)。

■ 设$\tilde{\mu}_Y$是μ_Y的另一个估计量，并假设$\hat{\mu}_Y$和$\tilde{\mu}_Y$都是无偏的，如果$\text{var}(\hat{\mu}_Y)<\text{var}(\tilde{\mu}_Y)$，那么称$\hat{\mu}_Y$比$\tilde{\mu}_Y$更有效(efficient)。

为了计算方便，这里假定观测期数n为偶数。$\widetilde{Y}$的均值为μ_Y，方差为$1.25\sigma_Y^2/n$（见练习3.7），因而，$\widetilde{Y}$是无偏的。由于当$n\to\infty$时，$\text{var}(\widetilde{Y})\to0$，因此，$\widetilde{Y}$又是一致的。不过，$\widetilde{Y}$的方差比$\overline{Y}$的方差大，所以$\overline{Y}$比$\widetilde{Y}$更有效。

估计量$\overline{Y}$，Y_1和$\widetilde{Y}$具有相同的数学结构：它们都是$Y_1,\cdots,Y_n$的加权平均。前面两段中的比较结果表明，Y_1和$\widetilde{Y}$具有比$\overline{Y}$更大的方差。实际上，这些结论反映了一个更一般的结果：$\overline{Y}$是$Y_1,\cdots,Y_n$的所有加权平均的无偏估计量中最有效的估计量。该结论在重要概念3.3中进行了阐述，并在第15章进行了证明。

$\overline{Y}$是μ_Y的最小二乘估计量　样本均值$\overline{Y}$为数据提供了最佳拟合，也就是说，观测值和$\overline{Y}$之间的均方差是所有可能估计量中最小的一个。



现在我们考虑使下式的值为最小时估计量 m 的求解问题。

$$\sum_{i=1}^{n}(Y_i - m)^2 \tag{3.2}$$

上式是估计量 m 和所有样本点之间总的离差或距离平方和的一种测度。由于 m 是 $E(Y)$ 的一个估计量,因此可以把它看做是 Y_i 值的预测。这样就可以把离差 $Y_i - m$ 看做是预测的误差。表达式(3.2)中离差的平方和可以被看做是预测误差的平方和。

使表达式(3.2)中 $Y_i - m$ 的离差平方和最小的估计量 m 被称为**最小二乘估计量**(least squares estimator)。可以用反复试验的方法来求解这个最小二乘问题:尝试用多个 m 值,直到你找到了一个值,这个值能够使表达式(3.2)的值尽可能地小。另一种方法就是像附录3.2中所描述的那样,利用代数或微积分方法来证明选择 $m=\overline{Y}$ 能够使表达式(3.2)中的离差平方和最小,也就是说,$\overline{Y}$ 是 μ_Y 的最小二乘估计量。

重要概念 3.3

$\overline{Y}$ 的有效性

设 $\hat{\mu}_Y$ 是 μ_Y 的一个估计量,同时 $\hat{\mu}_Y$ 是 $Y_1,\cdots,Y_n$ 的加权平均,即 $\hat{\mu}_Y=\frac{1}{n}\sum_{i=1}^{n}a_iY_i$,这里的 $a_1,\cdots,a_n$ 是非随机的常数。如果 $\hat{\mu}_Y$ 是无偏的,那么 $\mathrm{var}(\overline{Y})<\mathrm{var}(\hat{\mu}_Y)$,除非 $\hat{\mu}_Y=\overline{Y}$。也就是说,在 $Y_1,\cdots,Y_n$ 的所有加权平均的无偏估计量中,$\overline{Y}$ 是 μ_Y 最有效的估计量。

3.1.3 随机抽样的重要性

我们已经假定 $Y_1,\cdots,Y_n$ 是独立同分布的抽样样本,就像用简单随机抽样得到的样本一样。这个假定是非常重要的,因为非随机抽样会使 $\overline{Y}$ 是有偏的。假设为了估计全国的月度失业率,一个统计机构采用了一种抽样方案,即调查人员每月第二个星期三的上午10点钟访问坐在城市公园中处于劳动年龄的成年人。由于这时大多数在岗人员都在工作(而不会坐在公园里),用坐在公园中的人代表失业人员有些不太适当,因此,基于这个抽样方案的失业率的估计值将会是有偏的。这个偏差产生的原因在于这个抽样方案过度代表了或过度抽取了总体中的失业人员。虽然这个例子是虚构的,但是,"兰顿获胜"这个一般兴趣框中给出的由不完全随机抽样所引起的偏差的例子却是真实的。

设计一个样本选择方案,以偏差最小为准则,这一点是非常重要的。附录3.1论述了美国劳工统计局在实施美国当前人口调查(CPS)时的实际做法,该统计局就是用这种调查来估计美国月度失业率的。

一般兴趣框

兰顿获胜

在1936年的总统大选前夕,《文学报》发布的一项民意调查表明,阿尔夫·兰顿(Alf M. Landon)会以57%对43%的压倒性优势击败富兰克林·罗斯福(Franklin D. Roosevelt)。《文学报》报道的关于选举是压倒性优势这一结论是正确的,但是关于获胜者的预测却是错误的:罗斯福以59%对41%获胜!

《文学报》怎么会犯这么大的错误呢?《文学报》的样本是从电话记录和汽车登记文件中选取的,但在1936年,许多家庭还没有汽车或电话,那些比较富有的人更有可能是共和党人。由于电话调查并没有从总体中随机地抽样而是偏低地抽取了民主党人,因此估计量是



有偏的，从而令《文学报》犯了一个尴尬的错误。

你认为通过因特网进行调查会产生类似的偏差问题吗？

3.2 关于总体均值的假设检验

许多关于我们身边世界的假设都可以被归结为是或否的问题。美国刚毕业的大学生的平均每小时收入等于20美元吗？男女大学毕业生的平均收入相等吗？这两个问题体现了收入总体分布的特定假设。统计学面临的挑战就是根据样本证据来回答这些问题。本节描述了总体均值的检验假设（testing hypothesis）（如每小时收入的总体均值是否等于20美元）。关于两个总体的假设检验（如男女平均收入是否相同）将在第3.4节中介绍。

3.2.1 零假设和备择假设

统计假设检验的出发点，是设定一个待检验的假设，这个假设被称为**零假设**（null hypothesis）。假设检验要求利用数据来比较零假设和另一个假设，这个假设被称为**备择假设**（alternative hypothesis）。如果零假设不成立，那么备择假设就成立。

零假设是总体均值 $E(Y)$ 取一个特定的值，记为 $\mu_{Y,0}$。零假设用 H_0 表示，因此：

$$H_0: E(Y) = \mu_{Y,0} \tag{3.3}$$

例如，关于总体均值的推测——"大学毕业生每小时挣20美元"便构成了每小时收入总体分布的零假设。用数学语言表达，如果 Y 是随机选择的刚毕业大学生的每小时收入，那么零假设就是 $E(Y)=20$，即公式(3.3)中的 $\mu_{Y,0}=20$。

备择假设设定了当零假设不正确时的真实情况。备择假设最一般的形式是 $E(Y)\neq\mu_{Y,0}$，这被称为**双边备择假设**（two-sided alternative hypothesis），因为它允许 $E(Y)$ 小于或大于 $\mu_{Y,0}$。双边备择假设可写为：

$$H_1: E(Y) \neq \mu_{Y,0} \quad （双边备择假设） \tag{3.4}$$

单边备择假设也是可能的，相关的内容将在本节的后面进行讨论。

统计学家面对的问题是：利用随机选择的样本数据中的证据来判定，是接受零假设 H_0，还是拒绝零假设，进而接受备择假设 H_1。如果零假设被"接受"，这并不意味着统计学家宣布了零假设是真的，而是应该这样来理解，即在当前的证据下我们只好接受它，但是随着以后更多的证据被提供，零假设也可能被拒绝。因为这个原因，统计假设检验可被归结为：要么拒绝零假设，要么不能拒绝零假设。

3.2.2 p 值

在任意给定的样本中，样本均值 $\bar{Y}$ 很少会恰好等于所假设的值 $\mu_{Y,0}$。实际上，$\bar{Y}$ 和 $\mu_{Y,0}$ 之间的差异可能是由真实均值并不等于 $\mu_{Y,0}$（零假设是假的）所引起的，也可能是由虽然真实均值等于 $\mu_{Y,0}$（零假设是真的）但因随机抽样使得 $\bar{Y}$ 不等于 $\mu_{Y,0}$ 所引起的。要确切地区分这两种可能性是不可能的。虽然样本数据不能提供零假设的结论性证据，但是进行概率计算却是可能的，这种概率计算允许以解释抽样不确定性的方式来检验零假设。这个计算涉及用数据来计算零假设的 p 值。

p 值（p-value），又称为**显著性概率**（significance probability），是在假定零假设正确的情况下，你抽到了一个就像你从样本中实际计算出来的那个至少和零假设相背离的一个统



计量的概率。在我们眼前的例子中，p 值是在零假设下抽到的 $\overline{Y}$ 在其分布的尾部至少和实际计算的样本均值一样远的概率。

例如，假设在刚毕业的大学生的样本中，平均工资是22.24美元。p 值就是当零假设为真时，观测到一个 $\overline{Y}$ 的值至少和由纯随机抽样变化得到的观测值22.24美元一样不同于20美元（零假设下的总体均值）的概率。如果这个 p 值很小，比如说0.5%，那么在零假设为真的条件下抽到这个样本是不太可能的，因而，得出零假设确实不正确的结论是合理的。另一方面，如果这个 p 值很大，比如说40%，那么在零假设为真的条件下，观测的样本平均值22.24美元可能正好是由随机抽样变化所引起的，这是非常可能的，因而在这个概率意义下拒绝零假设的证据很弱，也就是说不拒绝零假设是合理的。

现在我们用数学语言来表达 p 值的定义。设 $\overline{Y}^{act}$ 为用我们手边的样本数据集实际计算的样本平均值，并设 $\Pr_{H_0}$ 表示在零假设下所计算的概率（即在假设 $E(Y_i)=\mu_{Y,0}$ 下所计算的概率），p 值为：

$$p\text{ 值}=\Pr_{H_0}[|\overline{Y}-\mu_{Y,0}|>|\overline{Y}^{act}-\mu_{Y,0}|] \tag{3.5}$$

即 p 值是零假设下在 $\overline{Y}$ 分布的尾部中超出 $|\overline{Y}^{act}-\mu_{Y,0}|$ 的面积。如果 p 值很大，那么观测值 $\overline{Y}^{act}$ 与零假设一致，但如果 p 值很小，$\overline{Y}^{act}$ 与零假设就不一致。

为了计算 p 值，必须要知道在零假设下 $\overline{Y}$ 的抽样分布。如2.6节中所讨论的，当样本容量很小时，这个分布是复杂的。不过，根据中心极限定理，当样本容量很大时，$\overline{Y}$ 的抽样分布可以非常好地被正态分布近似地表达。在零假设下，这个正态分布的均值是 $\mu_{Y,0}$，所以在零假设下 $\overline{Y}$ 服从分布 $N(\mu_{Y,0},\sigma_{\overline{Y}}^2)$，这里 $\sigma_{\overline{Y}}^2=\sigma_Y^2/n$。只要样本容量充分地大，这个大样本的正态近似使得在不需要知道 Y 的总体分布的情况下计算 p 值成为可能。但计算的细节依赖于 σ_Y^2 是已知还是未知。

3.2.3 σ_Y 已知时 p 值的计算

当 σ_Y 已知时，p 值的计算方法在图3—1中进行了总结。如果样本容量很大，那么，在零假设下 $\overline{Y}$ 的抽样分布是 $N(\mu_{Y,0},\sigma_{\overline{Y}}^2)$，这里 $\sigma_{\overline{Y}}^2=\sigma_Y^2/n$。因此，在零假设下，$\overline{Y}$ 的标准化形式 $(\overline{Y}-\mu_{Y,0})/\sigma_{\overline{Y}}$ 服从标准正态分布。p 值是在零假设下得到比 $\overline{Y}^{act}$ 更远离于 $\mu_{Y,0}$ 的 $\overline{Y}$ 的值的概率，或者说，得到 $(\overline{Y}-\mu_{Y,0})/\sigma_{\overline{Y}}$ 的绝对值大于 $(\overline{Y}^{act}-\mu_{Y,0})/\sigma_{\overline{Y}}$ 的绝对值的概率。这个概率就是图3—1中所显示的阴影部分的面积。用数学公式表示，图3—1中尾部阴影部分的概率（即 p 值）为：

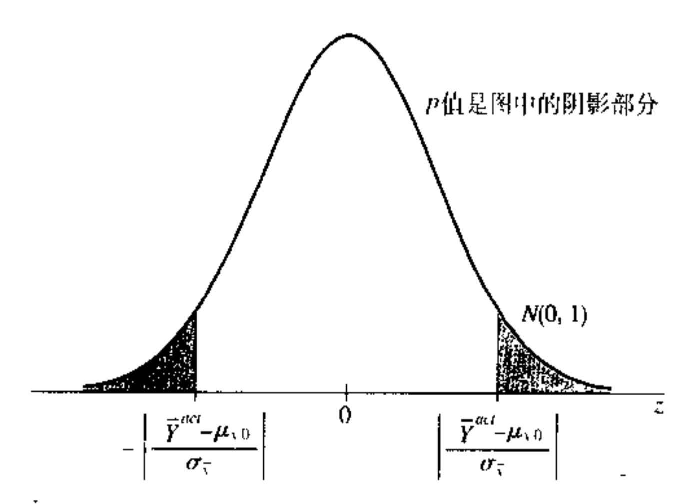
$$p\text{ 值}=\Pr_{H_0}\left(\left|\frac{\overline{Y}-\mu_{Y,0}}{\sigma_{\overline{Y}}}\right|>\left|\frac{\overline{Y}^{act}-\mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right)=2\Phi\left(-\left|\frac{\overline{Y}^{act}-\mu_{Y,0}}{\sigma_{\overline{Y}}}\right|\right) \tag{3.6}$$

其中，Φ 是标准正态累积分布函数。也就是说，p 值是标准正态分布在 $\pm(\overline{Y}^{act}-\mu_{Y,0})/\sigma_{\overline{Y}}$ 之外的尾部区域的面积。

公式(3.6)中的 p 值表达式依赖于总体分布的方差 σ_Y^2。在实践中，这个方差通常是未知的（一个例外是，当 Y_i 是二元变量时，它服从贝努里分布，这种情况下方差是由零假设决定的，见公式(2.7)）。由于一般在计算 p 值之前总是必须先估计出 σ_Y^2，因此，我们现在转向研究 σ_Y^2 的估计问题。

3.2.4 样本方差、样本标准差与标准误

样本方差 s_Y^2 是总体方差 σ_Y^2 的一个估计量；样本标准差 s_Y 是总体标准差 σ_Y 的一个估计



注：p 值是抽到的 $\overline{Y}$ 值至少和 $\overline{Y}^{act}$ 一样不同于 $\mu_{Y,0}$ 的概率。在大样本中，$\overline{Y}$ 在零假设下服从分布 $N(\mu_{Y,0},\sigma_{\overline{Y}}^2)$，所以，$(\overline{Y}-\mu_{Y,0})/\sigma_{\overline{Y}}$ 服从分布 $N(0,1)$。因此 p 值是标准正态分布在 $\pm|(\overline{Y}^{act}-\mu_{Y,0})/\sigma_{\overline{Y}}|$ 之外的尾部阴影部分的概率。

图3—1　p 值的计算

量；样本均值 $\overline{Y}$ 的标准误是 $\overline{Y}$ 的抽样分布标准差的一个估计量。

样本方差和标准差。**样本方差**（sample variance）s_Y^2 的计算公式为：

$$s_Y^2=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2 \tag{3.7}$$

样本标准差（sample standard deviation）s_Y 是样本方差的平方根。

样本方差的表达式与总体方差的表达式非常相似。总体方差 $E(Y-\mu_Y)^2$ 是总体分布中 $(Y-\mu_Y)^2$ 的平均值。同样，样本方差是 $(Y_i-\mu_Y)^2(i=1,\cdots,n)$ 的样本平均值，但有两个修正：一个是用 $\overline{Y}$ 代替了 μ_Y，另一个是用除数 $n-1$ 代替了 n。

第一个修正也即用 $\overline{Y}$ 代替 μ_Y，是因为 μ_Y 是未知的，所以必须估计，自然地，μ_Y 的估计量就是 $\overline{Y}$。第二个修正也即用除数 $n-1$ 代替 n，是由于用 $\overline{Y}$ 估计 μ_Y 时在 $(Y_i-\overline{Y})^2$ 中引入了一个很小的向下偏差。具体地讲，正如在练习3.11中所证明的，$E[(Y_i-\overline{Y})^2]=[(n-1)/n]\sigma_Y^2$，因而，$E\sum_{i=1}^{n}(Y_i-\overline{Y})^2=nE[(Y_i-\overline{Y})^2]=(n-1)\sigma_Y^2$。在公式(3.7)中用 $n-1$ 代替 n 作除数修正了这个小的向下偏差，因此 s_Y^2 是无偏的。

在公式(3.7)中，用 $n-1$ 作除数而不用 n 作除数，这被称为自由度（degrees of freedom）的修正。估计均值用去了部分信息，即用去了1个"自由度"，所以只剩下 $n-1$ 个自由度。

样本方差的一致性。样本方差是总体方差的一致性估计量，即：

$$s_Y^2\xrightarrow{p}\sigma_Y^2 \tag{3.8}$$

换句话说，当 n 很大时，样本方差以很高的概率逼近于总体方差。

在 $Y_1,\cdots,Y_n$ 为独立同分布以及 Y_i 具有有限的四阶矩，即 $E(Y_i^4)<\infty$ 的假设下，附录3.3证明了表达式(3.8)中的结论。直观地分析，s_Y^2 是一致性的原因在于它是样本的均值，所以 s_Y^2 服从大数定律。但由于 s_Y^2 服从大数定律（见重要概念2.6），因此，$(Y_i-\mu_Y)^2$ 必须具有有限方差，反过来，这就意味着 $E(Y_i^4)$ 一定是有限的，也就是说，Y_i 一定具有有限的四阶矩。

样本均值的概率是3.9%。

当Y服从正态分布时t统计量的分布　当总体服从正态分布时，t统计量服从自由度为$n-1$的学生t分布(见2.4节)，因此在这个特殊情况下，对于任意的样本容量n，不依赖中心极限定理就能精确地计算p值。由于学生t分布尾部的面积要大于正态分布的尾部面积，因此，利用学生t分布计算的p值要比用正态分布计算的p值略大一些。

虽然一些统计软件利用学生t分布来计算p值，但是本书不使用这个分布计算p值，原因有两个。第一，t统计量只有在总体服从正态分布的条件下才服从学生t分布，而学生t分布常常与经济数据的实际分布较差地近似。因此，当Y服从正态分布时，这个分布精确成立的优势被它不太适用的事实给否定了。第二，如果样本容量适度，学生t分布和正态分布的差别非常小，而如果样本容量很大，这个差别可以忽略。对于$n>15$，用t分布和正态分布所计算的p值的差不会超过0.01，而对于$n>80$，它们的差绝不会超过0.002。在现代应用中，以及在本书所有的应用中，样本容量都是数以百千计的，因此，样本容量大得足以忽略学生t分布和正态分布之间的差异。

3.2.7　事先设定显著性水平条件下的假设检验

假设已经决定，如果p值小于5%，那么就拒绝零假设。由于在正态分布的尾部超过± 1.96部分的面积是5%，因此这给出了一条简单的规则：

$$\text{如果}|t^{act}|>1.96\text{，那么就拒绝零假设}H_0 \tag{3.15}$$

也就是说，如果由样本所计算的t统计量的绝对值大于1.96，那么就拒绝零假设。如果n足够大，那么在零假设下t统计量服从$N(0,1)$分布。因此，错误地拒绝零假设(当零假设实际为真时却拒绝了零假设)的概率是5%。

这个检验统计假设的概念框架包含一些专业术语，这些术语在重要概念3.5中做了总结。在表达式(3.15)中检验的显著性水平是5%，这个双边检验的临界值是1.96，拒绝域是t统计量在± 1.96以外的值。如果检验在5%的显著性水平下拒绝零假设，那么就可以说，在5%的显著性水平下总体均值在统计上显著地不同于$\mu_{Y,0}$。

重要概念3.5

假设检验的术语

在零假设下，一个统计假设检验事先设定的拒绝概率就是该检验的显著性水平(significance level)。检验统计量的临界值(critical value)，是指在给定的显著性水平下检验恰好拒绝了零假设的那个统计量的值。检验拒绝零假设时该检验统计量的值的集合称为拒绝域(rejection region)，不能拒绝零假设的检验统计量的值的集合称为接受域(acceptance region)。当零假设为真时，检验结果实际上却不正确地拒绝了零假设的概率，这称为检验的级效(size，又称检验的级。译者注)；当备择假设为真时，检验结果正确地拒绝了零假设的概率，这称为检验的功效(power，又称检验的势，译者注)。

p值是在假定零假设为真的情况下，由随机抽样变化得到的检验统计量至少与实际观测到的统计量一样背离零假设的概率。同理，p值是拒绝零假设的最小显著性水平。

利用事先设定的显著性水平检验假设，并不需要计算p值。在前面的检验假设即检验"刚毕业大学生的平均收入为20美元/小时"的例子中，t统计量为2.06，它超过了1.96，所以零假设在5%的显著性水平下遭到了拒绝。虽然以5%的显著性水平比较容易进行假设



检验,但是仅仅报告在事先给定的显著性水平下是否拒绝零假设所传达的信息比报告 p 值所传达的信息要少。

实践中应采用多大的显著性水平?在很多情况下,统计学家和经济计量学家使用5%的显著性水平。如果你要在5%的显著性水平下检验多个统计假设,那么平均每20个检验中,你就会错误地拒绝一个零假设。有时设定更加保守的显著性水平可能是适宜的。例如,法律案件有时涉及统计证据,零假设可能是被告人无罪,那么,人们可能会非常希望得到确定性的结论,比如拒绝零假设(有罪的结论)并不是随机样本变化的结果。因此在一些法律案件中,经常用1%甚至0.1%来设定显著性水平以避免此类错误的发生。同样,如果政府机构正考虑批准一种新药的销售,非常保守的标准可能是合适的,这样可以保证消费者在市场上买到的药品确实有效。

使用非常低的显著性水平,也就是变得非常保守,这是需要付出代价的:当零假设为假时,显著性水平越小,临界值越大,拒绝零假设就越困难。实际上,最保守的做法就是永远不拒绝零假设,但如果你真的那样做的话,你永远也不需要调查任何的统计证据,因为你绝不会改变你的看法!显著性水平越低,检验的功效就越低。许多经济和政策方面的应用要求的保守程度比法律案件要宽松一些,因此通常认为5%的显著性水平是个合理的折中的方案。

关于总体均值双边备择假设的假设检验的程序,在重要概念3.6中做了总结。

重要概念3.6

检验假设 $E(Y)=\mu_{Y,0}$ 与对应的备择假设 $E(Y)\neq\mu_{Y,0}$

1. 计算 $\bar{Y}$ 的标准误 $SE(\bar{Y})$(公式(3.14))。
2. 计算 t 统计量(公式(3.10))。
3. 计算 p 值(公式(3.13))。如果 p 值小于0.05(同理,如果 $|t^{act}|>1.96$),那么在5%的显著性水平下拒绝零假设。

3.2.8 单边备择假设检验

在某些情况下,备择假设可能是均值大于 $\mu_{Y,0}$。例如,人们希望教育在劳动力市场上有积极作用,因此,相对于大学毕业生和非大学毕业生收人相同的零假设而言,备择假设不再是他们的收入有差别,而是大学毕业生要比非大学毕业生的收人高。这种备择假设被称为单边备择假设(one-sided alternative hypothesis),可写为:

$$H_1:E(Y)>\mu_{Y,0} \quad (单边备择假设) \tag{3.16}$$

对单边备择假设而言,计算 p 值和假设检验的一般方法与双边备择假设的方法相同,只是有一点修改,即 t 统计量只有大的正的值才会拒绝零假设,而不是大的绝对值。具体来讲,为了检验不等式(3.16)中的单边假设,需要构造公式(3.10)中的 t 统计量。p 值是在标准正态分布下所计算的 t 统计量右侧的面积。也就是说,基于 t 统计量分布的标准正态分布 $N(0,1)$ 近似原理,p 值为:

$$p\text{ 值}=\Pr_{H_0}(Z>t^{act})=1-\Phi(t^{act}) \tag{3.17}$$

在5%的显著性水平下单边检验 $N(0,1)$ 分布的临界值是1.645。该检验的拒绝域就是所有大于1.645的 t 统计量的值。

不等式(3.16)中的单边假设关注 μ_Y 大于 $\mu_{Y,0}$ 的值。如果反过来,备择假设是 $E(Y)<$



总体均值置信区间的覆盖概率(coverage probability)、是指通过重复抽样计算的该区间包含真实总体均值的概率。



重要概念 3.7

总体均值的置信区间

μ_Y 的 95% 的双边置信区间是这样构造的一个区间，在各种应用中有 95% 的应用场合包含有 μ_Y 的真实值。当样本容量 n 很大时，μ_Y 的 95%，90% 和 99% 的置信区间分别是：

μ_Y 的 95% 的置信区间 $= \{\overline{Y} \pm 1.96SE(\overline{Y})\}$

μ_Y 的 90% 的置信区间 $= \{\overline{Y} \pm 1.64SE(\overline{Y})\}$

μ_Y 的 99% 的置信区间 $= \{\overline{Y} \pm 2.57SE(\overline{Y})\}$

3.4 不同总体均值的比较

平均来说，刚毕业的男女大学生的收入相同吗？这个问题涉及比较两个不同总体分布的均值。本节归纳了如何检验两个不同总体均值差的假设以及如何构造其置信区间的问题。

3.4.1 两个均值的差的假设检验

设 μ_w 为刚毕业的女大学生总体平均每小时收入，并设 μ_m 为刚毕业的男大学生总体平均每小时收入，考虑零假设：这两个总体的收入差为一个确定的数，比如说 d_0，那么零假设和双边备择假设为：

$$H_0:\mu_m - \mu_w = d_0 \quad H_1:\mu_m - \mu_w \neq d_0 \tag{3.18}$$

在这两个总体中，"男女具有相同收人"这一零假设，与表达式(3.18)中当 $d_0 = 0$ 时的零假设 H_0 是对应的。

由于这些总体的均值是未知的，因此必须使用男女的样本数据进行估计。假设我们拥有从总体中随机抽取的 n_m 名男大学生样本和 n_w 名女大学生样本。设男、女大学生样本的平均年收入分别为 $\overline{Y}_m$ 和 $\overline{Y}_w$，那么 $\mu_m - \mu_w$ 的估计量就是 $\overline{Y}_m - \overline{Y}_w$。

利用 $\overline{Y}_m - \overline{Y}_w$ 检验零假设 $\mu_m - \mu_w = d_0$，我们需要知道 $\overline{Y}_m - \overline{Y}_w$ 的分布。根据中心极限定理，我们知道 $\overline{Y}_m$ 近似地服从分布 $N(\mu_m, \sigma_m^2/n_m)$，其中 σ_m^2 是男生收入的总体方差。同理，$\overline{Y}_w$ 近似服从分布 $N(\mu_w, \sigma_w^2/n_w)$，其中 σ_w^2 是女生收入的总体方差。另外，由 2.4 节可知，两个正态随机变量的加权平均本身也服从正态分布。因为 $\overline{Y}_m$ 和 $\overline{Y}_w$ 是由两个不同的随机选择的样本构造出来的，所以它们是独立随机变量。因此，$\overline{Y}_m - \overline{Y}_w$ 服从分布 $N[\mu_m - \mu_w, (\sigma_m^2/n_m) + (\sigma_w^2/n_w)]$。

如果 σ_m^2 和 σ_w^2 是已知的，那么这个近似正态分布可被用来计算检验零假设 $\mu_m - \mu_w = d_0$ 的 p 值。然而，实际上这些总体方差通常是未知的，因此必须对它们进行估计。如前所述，可以用样本方差 s_m^2 和 s_w^2 来估计它们，这里 s_m^2 的定义在公式(3.7)中已经给出了，不同的是这个统计量是只对样本中的男生进行计算得出的，女生 s_w^2 的定义方式与之相同。因此，$\overline{Y}_m - \overline{Y}_w$ 的标准误为：

$$SE(\overline{Y}_m - \overline{Y}_w) = \sqrt{\frac{s_m^2}{n_m} + \frac{s_w^2}{n_w}} \tag{3.19}$$

构造检验零假设的 t 统计量与检验单个总体均值假设的 t 统计量类似，即通过将估计量 $\overline{Y}_m-\overline{Y}_w$ 减去 $\mu_m-\mu_w$ 的零假设值，再除以 $\overline{Y}_m-\overline{Y}_w$ 的标准误，即：

$$t=\frac{(\overline{Y}_m-\overline{Y}_w)-d_0}{SE(\overline{Y}_m-\overline{Y}_w)} \quad \text{（比较两个均值的 } t \text{ 统计量）} \tag{3.20}$$

如果 n_m 和 n_w 都很大，那么这个 t 统计量服从标准正态分布。①

因为当 n_m 和 n_w 很大时，公式(3.20)中的 t 统计量在零假设下服从标准正态分布，所以，双边检验 p 值的计算方法与单个总体中 p 值的计算方法完全相同，即 p 值可以用公式(3.13)来计算。

要进行一个事先设定显著性水平的检验，只需计算出公式(3.20)中的 t 统计量，并将它和相应的临界值做比较。例如，如果 t 统计量的绝对值超过了1.96，那么可在5%的显著性水平下拒绝零假设。

如果备择假设是单边的而不是双边的，即如果备择假设是 $\mu_m-\mu_w>d_0$，那么，检验应该按照3.2节中所总结的那样进行调整。利用公式(3.17)计算 p 值，当 $t>1.65$ 时，检验在5%的显著性水平下拒绝零假设。

3.4.2 两个总体均值差的置信区间

在3.3节中所总结的构造置信区间的方法可推广到构造均值之差 $d=\mu_m-\mu_w$ 的置信区间。因为如果 $|t|>1.96$，那么在5%的显著性水平下拒绝假设值 d_0，所以如果 $|t|\leq 1.96$，那么 d_0 将会在置信集中。但是，$|t|\leq 1.96$ 意味着估计偏差 $\overline{Y}_m-\overline{Y}_w$ 距离 d_0 不超过1.96个标准误。因此，d 的95%的双边置信区间由 $\overline{Y}_m-\overline{Y}_w$ 在 ± 1.96 个标准误的范围内的那些 d 值组成：

$$d=\mu_m-\mu_w \text{ 的95\%的置信区间}=(\overline{Y}_m-\overline{Y}_w)\pm 1.96SE(\overline{Y}_m-\overline{Y}_w) \tag{3.21}$$

熟悉了这些公式，我们现在开始对美国大学毕业生不同性别的收入差异进行实证研究。

3.5 美国男女大学毕业生的收入问题

历史上男性比女性更易于找到好报酬的工作。但是，社会规范的变化和禁止性别歧视法律的实施已使男女在现代工作场所中具有平等的角色——至少在理论上是这样。但如果存在性别收入差异的话，受到良好教育的青年男女之间的收入差异实际上是多少呢？

表3—1给出了美国年龄在25~34岁之间的大学毕业的全职劳动者每小时收入的估计值。表3—1中的统计数据是根据当前人口普查(CPS)的部分数据计算得到的，附录3.1中描述了CPS。利用消费价格指数②将所有的收入调整为1998年的美元价值水平，也即对通

---

① 如果两个总体的方差相等(即 $\sigma_m^2=\sigma_w^2=\sigma^2$)，那么 $\overline{Y}_m-\overline{Y}_w$ 服从分布 $N(\mu_m-\mu_w,[(1/n_m)+(1/n_w)]\sigma^2)$。在这种特殊情况下，可以使用所谓的 σ^2 的合并方差估计量(pooled variance estimator)：

$$s_{pooled}^2=\frac{1}{n_m+n_w-2}\left[\sum_{i=1}^{n_m}(Y_i-\overline{Y}_m)^2+\sum_{j=1}^{n_w}(Y_j-\overline{Y}_w)^2\right]$$

这个表达式中的第一个求和项是对男生观测值而言的，而第二个求和项是对女生观测值而言的。如果总体方差不同，那么合并方差估计量则是有偏的，而且是不一致的。因此，除非有很好的理由相信总体方差是相同的，否则在实际中不应该使用合并方差估计量。

② 由于通货膨胀，1992年的1美元要比1998年的1美元更值钱，也就是说，1992年的1美元能够购买到的商品和服务比1998年的1美元多。因此，如果不进行通货膨胀调整，那么1992年的收入和1998年的收入是不可直接对比的。进行这种调整的一种方法就是利用消费者价格指数(CPI)，它是由劳工统计局设计的对消费者提供商品和服务的"市场篮子"价格的测度。从1992年到1998年的6年间，CPI市场篮子的价格上涨了16.2%，也就是说，在1992年花100美元购买到的商品和服务的CPI篮子在1998年要花费116.20美元。为了使表3—1中1992年和1998年的收入可进行比较，1992年的收入要与总体CPI价格指数相乘，即把1992年的收入乘以1.162，将它们转化为"1998年的美元"。

货膨胀进行了调整。

表 3—1　按 1998 年美元计算的年龄在 25～34 岁之间美国大学毕业职工的每小时收入
（来自于当前人口普查的部分统计数据）

年份	男			女			男女差异		
	$\bar{Y}_m$	s_m	n_m	$\bar{Y}_w$	s_w	n_w	$\bar{Y}_m-\bar{Y}_w$	$SE(\bar{Y}_m-\bar{Y}_w)$	d 的 95% 的置信区间
1992	17.57	7.50	1 591	15.22	5.97	1 371	2.35**	0.25	1.87～2.84
1994	16.93	7.39	1 598	15.01	6.41	1 358	1.92**	0.25	1.42～2.42
1996	16.88	7.29	1 374	14.42	6.07	1 235	2.46**	0.26	1.94～2.97
1998	17.94	7.86	1 393	15.49	6.80	1 210	2.45**	0.29	1.89～3.02

注：这些估计值是利用所示年份 CPS 的年龄在 25～34 岁之间的全职工人的数据计算的。在*5% 或**1% 的显著性水平下，这个差显著地异于零。

表 3—1 的前三列给出了男性收入的描述性信息；接下来的三列给出了女性收入的描述性信息；最后三列给出了男女收入之差的信息。例如，1999 年 3 月实施的当前人口普查（CPS）共调查了 64 000 个家庭，其中包括 1 393 个年龄在 25～34 岁之间拥有大学学位的全职男职工。这 1 393 个男性在 1998 年平均每小时的收入是 17.94 美元，这些收入的标准差是 7.86 美元。在那次调查中，1 210 个女性的平均每小时收入是 15.49 美元，标准差是 6.80 美元。工资差为 17.94 − 15.49 = 2.45（美元/小时），标准误是 $\sqrt{(7.86^2/1\ 393)+(6.80^2/1\ 210)}=0.29$（美元）。所以，检验工资差额为 0 的 t 统计量是 $(2.45-0)/0.29=8.45$，它超过了 1% 显著性水平下的双边检验临界值 2.58，所以它在 1% 的水平下（实际上，它在 0.01% 的水平下也是显著的）是显著的。这个收入差的 95% 的置信区间是 $2.45\pm1.96\times0.29$ 美元 =（1.89 美元，3.02 美元）。也就是说，在 95% 的置信水平下，我们估计这两个总体之间的工资差在 1.89 美元和 3.02 美元之间。

男女工资的差额比较大。根据表 3—1 中的估计值，在 1998 年，女性的每小时收入比男性少 14%（2.45/17.94）。此外，这个差额在 20 世纪 90 年代没有发生多大的变化，这个估计差额不太可能仅仅是由人为的抽样误差造成，因为 1.89 美元仅仅是包含在 1998 年 95% 的置信区间里收入差额的最小值。

这个统计分析表明了在每小时平均收入上存在“性别差异”，但它并没有说明这个差异的来源或原因。这个差异是劳动力市场中性别歧视的结果，还是在工作技能和经验方面男性和女性之间有差别进而导致收入的不同？要解决这些问题，我们需要借助于多元回归分析，这也是本书第 2 部分的主题。不过，首先我们必须复习散点图、样本协方差和样本相关系数。

3.6 散点图、样本协方差和样本相关系数

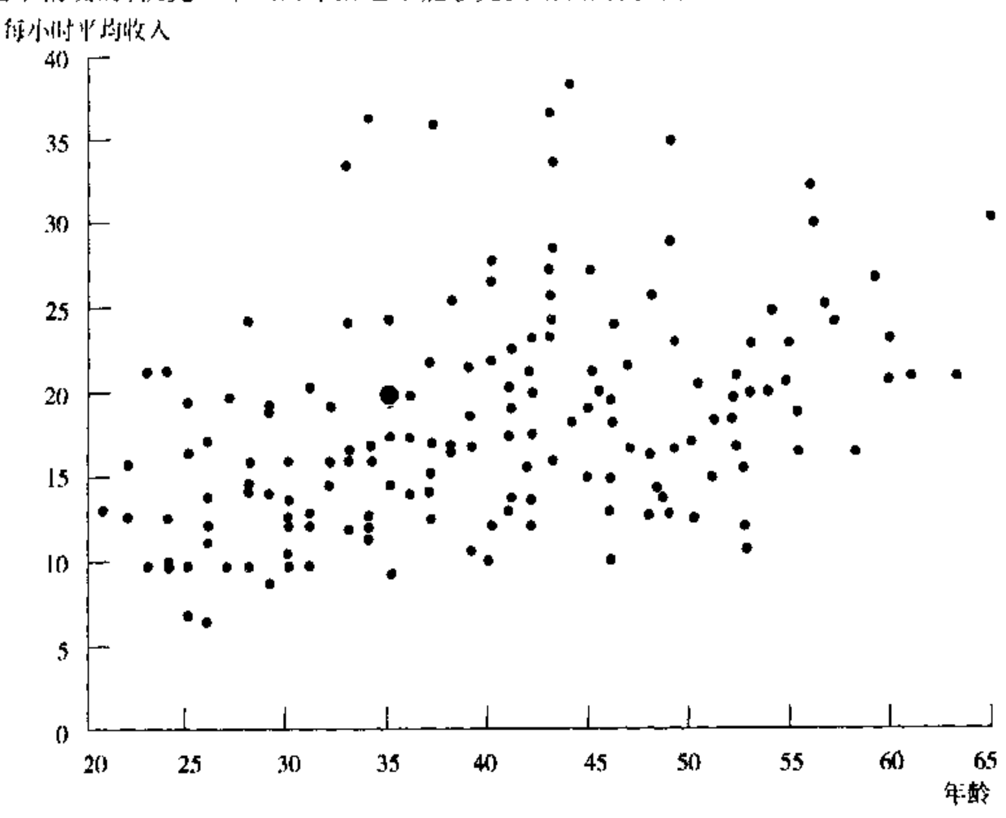
年龄和收入之间有什么关系？像许多其他问题一样，这个问题将一个变量 X（年龄）与另一个变量 Y（收入）联系在一起。本节回顾总结了反映变量之间关系的三种方法：散点图、样本协方差和样本相关系数。

3.6.1 散点图

散点图（scatter plot）是关于 X_i 和 Y_i 的 n 个观测值的图形，其中每个观测值是用点



(X_i, Y_i)来表示的。例如,图3—2是来自1999年3月CPS的在通讯业中工作但没有大学学历的184名技工样本的年龄(X)和每小时收入(Y)的散点图。图3—2中的每一点都对应于一对观测值(X, Y)。例如,这个样本中有一个35岁的技工,每小时收入为19.61美元,这个技工的年龄和收入由图3—2中较大的彩色点表示。散点图显示了这个样本中年龄和收入之间的正相关关系:年龄较大的通讯技工趋向于比年龄较小的技工挣得更多。但是,这个关系是不精确的,仅凭一个人的年龄还不能够完全地预测他的收入。



注:图中的每一个点代表了184名工人样本中每一名工人的年龄和平均收入。图中较大的彩色点对应于平均收入为19.61美元的一名35岁工人。数据来自于1999年3月CPS中没有大学学历的通讯业技工。

图3—2 平均每小时收入与年龄的散点图

3.6.2 样本协方差和样本相关性

协方差和相关性(相关系数)作为随机变量X和Y的联合概率分布的两个性质,我们在2.3节中已经介绍过了。因为总体分布是未知的,所以实际上我们不知道总体的协方差或相关系数。不过,通过抽取总体中n个个体的随机样本并收集数据(X_i, Y_i), $i = 1, \cdots, n$,我们便能够估计出总体的协方差和相关系数。

样本协方差和相关系数是总体协方差和相关系数的估计量。和本章前面讨论的估计量一样,它们是通过用样本均值代替总体均值(期望)来计算的。样本协方差(sample covariance),用s_{XY}表示,其计算公式为:

$$s_{XY} = \frac{1}{n-1} \sum_{i=1}^{n} (X_i - \bar{X})(Y_i - \bar{Y}) \tag{3.22}$$

和样本方差一样,公式(3.22)中的平均数是通过除以$n-1$,而不是除以n计算得到的。产



生这个差异的原因是因为使用了$\overline{X}$和$\overline{Y}$来估计它们各自的总体均值。当n很大时，除数是n还是$n-1$的差别并不大。



样本相关系数(sample correlation coefficient)，或称样本相关性(sample correlation)是样本协方差与样本标准差之比，用r_{XY}表示，其计算公式为：

$$r_{XY}=\frac{s_{XY}}{s_X s_Y} \tag{3.23}$$

样本相关系数测度n个观测值的样本中X和Y之间线性相关的强度。和总体相关系数一样，样本相关系数也是没有单位的，其值介于-1和1之间，即$|r_{XY}|\leqslant 1$。

对于所有的i，当$X_i=Y_i$时样本相关系数等于1；对于所有的i，当$X_i=-Y_i$时样本相关系数等于-1。更一般地说，如果散点图是一条直线，那么相关系数为±1。如果该直线向上倾斜，则在X和Y之间存在正的相关性且相关系数为1；如果该直线向下倾斜，则X和Y之间存在负的相关性且相关系数为-1。散点图越接近于直线，相关系数越接近于±1。高的相关系数并不意味着直线的斜率很陡峭，而是说明散点图中的点非常趋近于一条直线。

3.6.3 样本协方差和样本相关系数的一致性

和样本方差一样，样本协方差也是一致的，即：

$$s_{XY}\xrightarrow{p}\sigma_{XY} \tag{3.24}$$

换句话说，在大样本中，样本协方差以很高的概率逼近于总体协方差。

在(X_i,Y_i)是独立同分布的且X_i和Y_i具有有限的四阶矩的假设下，表达式(3.24)的证明类似于附录3.3中样本协方差一致性的证明，这个证明留作练习(见练习15.2)。

由于样本方差和样本协方差都是一致的，因此样本相关系数也是一致的，即$r_{XY}\xrightarrow{p}\mathrm{corr}(X_i,Y_i)$。

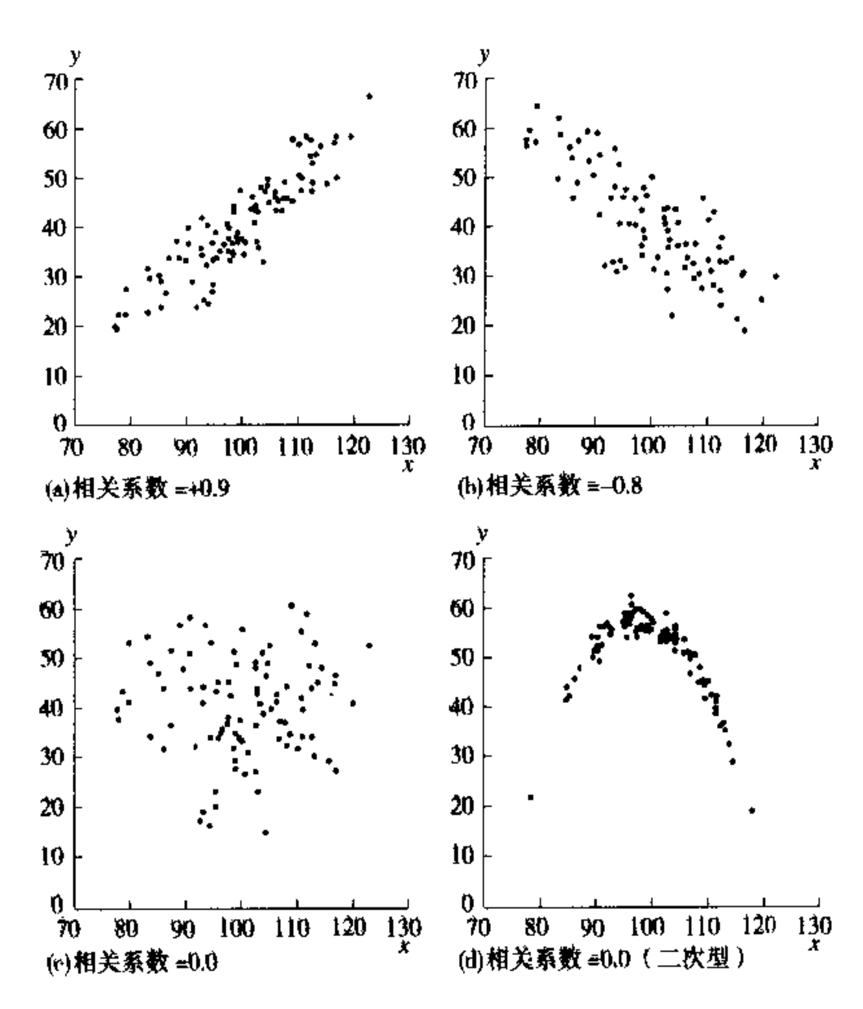
例子 作为一个例子，考虑图3—2中关于年龄和平均收入的数据。这184名工人的年龄的样本标准差为$s_A=10.49$年，收入的样本标准差为$s_E=6.44$美元/小时，年龄和收入之间的协方差为$s_{AE}=24.29$(单位是年×美元/小时，不容易解释)。因此，相关系数为$r_{AE}=24.29/(10.49\times6.44)=0.36$，或表示为36%。0.36的相关系数意味着在年龄和收入之间存在正向关系，但如散点图所揭露的，这个关系远不够完善。

为了证明相关系数不依赖于测度单位，假定收入用美分作单位，在这种情况下，收入的样本标准差为644美分/小时，年龄和收入之间的协方差为2 429(单位是年×美分/小时)，那么相关系数为$2\,429/(10.49\times644)=0.36$，或表示为36%。

图3—3给出了另外一些散点图和相关性的例子。图3—3(a)显示了这些变量之间很强的正线性关系，样本相关系数为0.9。图3—3(b)显示了样本相关系数为-0.8的较强的负线性关系。图3—3(c)显示了没有明显关系的散点图，样本相关系数为0。图3—3(d)显示了两个具有明显关系的变量：随着X的增加，Y先增大后减小。虽然X和Y之间存在可识别的关系，但是它们的样本相关系数却为0，原因就是对这些数据而言，小的Y值同大的和小的X值都有关。

最后的这个例子强调了重要的一点：相关系数是线性关系的一种测度。图3—3(d)中X和Y之间虽然存在关系，但不是线性的。





注：图 3—3(a) 和图 3—3(b) 中的散点图显示了 X 与 Y 之间很强的线性关系，在图 3—3(c) 中，X 独立于 Y，因此这两个变量不相关。在图 3—3(d) 中，即使这两个变量之间存在非线性关系，它们仍然是不相关的。

图 3—3　四个假设数据集的散点图

总结

1. 样本均值 $\overline{Y}$ 是总体均值 μ_Y 的一个估计量。当 $Y_1,\cdots,Y_n$ 独立同分布时：

a. $\overline{Y}$ 的抽样分布的均值为 μ_Y，方差为 $\sigma_{\overline{Y}}^2=\sigma_Y^2/n$；

b. $\overline{Y}$ 是无偏的；

c. 根据大数定律，$\overline{Y}$ 是一致的；

d. 根据中心极限定理，当样本容量很大时，$\overline{Y}$ 具有近似的正态抽样分布。

2. t 统计量用来检验总体均值取某一特定值的零假设。如果 n 很大，当零假设为真时，t 统计量具有标准正态抽样分布。

3. t 统计量可用来计算与零假设有关的 p 值。小的 p 值是零假设为假的证据。

4. μ_Y 的 95% 的置信区间是一个构造区间，因此，在 95% 的重复抽样中，它包含 μ_Y 的真值。

5. 两个总体均值差异的假设检验和置信区间在概念上类似于单个总体均值的假设检验和置信区间。

6. 样本相关系数是总体相关系数的一个估计量，它测度了两个变量之间的线性关系，即它们的散点图近似于直线的程度如何。



重要术语

估计量　估计值　偏差、一致性和有效性　最小二乘估计量　假设检验　零假设和备择假设　双边备择假设　p 值　样本方差　样本标准差　自由度　t 统计量　估计量的标准误　检验统计量　显著性水平　临界值　拒绝域　检验的级效　检验的功效　单边备择假设　置信集　置信水平　置信区间　覆盖概率　两个均值之差的检验　接受域　散点图　样本协方差和样本相关系数

复习概念

3.1　请解释样本均值 $\bar{Y}$ 和总体均值之间的差别。

3.2　请解释估计量和估计值之间的差别，并分别举一个例子。

3.3　已知总体分布的均值等于 10，方差等于 16。当：(a) $n=10$，(b) $n=100$，(c) $n=1\,000$时，确定从这个总体中抽取的一个独立同分布样本的 $\bar{Y}$ 的均值和方差，请将你的答案与大数定律联系起来。

3.4　中心极限定理在统计假设检验中起什么作用？在构造置信区间中又起什么作用？

3.5　零假设和备择假设的区别是什么？级效、显著性水平和功效的区别是什么？单边备择假设和双边备择假设的区别是什么？

3.6　为什么置信区间包含的信息要比单个的假设检验多？

3.7　当总体相关系数为：(a)1.0；(b) −1.0；(c)0.9；(d) −0.5；(e)0.0 时，画出样本容量为 10 的两个随机变量样本的散点图。

练习

3.1　在一个 $\mu_Y=100, \sigma_Y^2=43$ 的总体中，运用中心极限定理回答下列问题：

*a. 在一个容量 $n=100$ 的随机样本中，求 $\Pr(\bar{Y}<101)$；

b. 在一个容量 $n=64$ 的随机样本中，求 $\Pr(101<\bar{Y}<103)$；

c. 在一个容量 $n=165$ 的随机样本中，求 $\Pr(\bar{Y}>98)$。

3.2　设 Y 是个贝努里随机变量，其成功概率 $\Pr(Y=1)=p$，设从这个分布中抽取的 $Y_1, \cdots, Y_n$ 是独立同分布的，$\hat{p}$ 为这个样本的成功概率。

a. 证明：$\hat{p}=\bar{Y}$；

b. 证明：$\hat{p}$ 是 p 的无偏估计量；

c. 证明：$\text{var}(\hat{p})=p(1-p)/n$。

3.3　在对 400 名可能成为选民的一次调查中，有 215 人表示要支持在任者，而 185 人则表示要支持挑战者。设 p 代表在调查时希望支持在任者的所有可能选民的比例，用 $\hat{p}$ 表示被调查对象中支持在任者的选民比例。

*a. 利用调查结果估计 p。

*b. 利用 $\hat{p}$ 的方差估计量 $\hat{p}(1-\hat{p})/n$ 计算你的估计量的标准误。

$E(Y)$的最小二乘估计量。

微积分证明方法。为了使预测误差平方和最小化,我们对它求导,并令其导数为零。

$$\frac{d}{dm}\sum_{i=1}^{n}(Y_i-m)^2=-2\sum_{i=1}^{n}(Y_i-m)=-2\sum_{i=1}^{n}Y_i+2nm=0 \tag{3.25}$$

解最后一个方程中的 m,其结果表明,当 $m=\overline{Y}$ 时,会使 $\sum_{i=1}^{n}(Y_i-m)^2$ 最小化。

非微积分证明方法。这个方法就是证明最小二乘估计量和 $\overline{Y}$ 之差必为 0,由此得出 $\overline{Y}$ 是最小二乘估计量。设 $d=\overline{Y}-m$,则 $m=\overline{Y}-d$,那么$(Y_i-m)^2=[Y_i-(\overline{Y}-d)]^2=[(Y_i-\overline{Y})+d]^2=(Y_i-\overline{Y})^2+2d(Y_i-\overline{Y})+d^2$。因此,预测误差平方和(见表达式(3.2))为:

$$\sum_{i=1}^{n}(Y_i-m)^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+2d\sum_{i=1}^{n}(Y_i-\overline{Y})+nd^2=\sum_{i=1}^{n}(Y_i-\overline{Y})^2+nd^2 \tag{3.26}$$

其中,第二个等式利用了 $\sum_{i=1}^{n}(Y_i-\overline{Y})=0$ 这一事实。因为等式(3.26)的最后一个等式中的两项都是非负的,且前一项不依赖于 d,所以通过选择 d 使第二项 nd^2 尽可能地小,就能使 $\sum_{i=1}^{n}(Y_i-m)^2$ 最小化。通过设 $d=0$ 即 $m=\overline{Y}$ 就能做到这一点,因此,$\overline{Y}$ 是 $E(Y)$的最小二乘估计量。

附录 3.3 样本方差是一致性估计量的证明

本附录利用大数定律证明,当 $Y_1,\cdots,Y_n$ 是独立同分布的且 $E(Y_i^4)<+\infty$ 时,样本方差 s_Y^2 是总体方差 σ_Y^2 的一致性估计量,见表达式(3.8)中的陈述。

首先,加上和减去 μ_Y,则$(Y_i-\overline{Y})^2=[(Y_i-\mu_Y)-(\overline{Y}-\mu_Y)]^2=(Y_i-\mu_Y)^2-2(Y_i-\mu_Y)(\overline{Y}-\mu_Y)+(\overline{Y}-\mu_Y)^2$,然后,将表达式$(Y_i-\overline{Y})^2$ 代入到 s_Y^2 的定义(见公式(3.7))中,得到:

$$\begin{aligned} s_Y^2&=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\\ &=\frac{1}{n-1}\sum_{i=1}^{n}(Y_i-\mu_Y)^2-\frac{2}{n-1}\sum_{i=1}^{n}(Y_i-\mu_Y)(\overline{Y}-\mu_Y)+\frac{1}{n-1}\sum_{i=1}^{n}(\overline{Y}-\mu_Y)^2\\ &=\left(\frac{n}{n-1}\right)\left[\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)^2\right]-\left(\frac{n}{n-1}\right)(\overline{Y}-\mu_Y)^2 \end{aligned} \tag{3.27}$$

其中,最后一个等式是根据 $\overline{Y}$ 的定义(它意味着 $\sum_{i=1}^{n}(Y_i-\mu_Y)=n(\overline{Y}-\mu_Y)$)再合并同类项得到的。

现在将大数定律应用到公式(3.27)最后一行的两项中。定义 $W_i=(Y_i-\mu_Y)^2$,现有$E(W_i)=\sigma_Y^2$(根据方差的定义)。由于随机变量 $Y_1,\cdots,Y_n$ 是独立同分布的,因此,随机变量 $W_1,\cdots,W_n$ 也是独立同分布的。此外,根据假设 $E(Y_i^4)<+\infty$,有 $E(W_i^2)=E[(Y_i-\mu_Y)^4]<+\infty$。因此,$W_1,\cdots,W_n$ 是独立同分布的,且 $\text{var}(W_i)<+\infty$。所以 $\overline{W}$ 满足重要概念 2.6 中大数定律的条件,且 $\overline{W}\xrightarrow{p}E(W_i)$。但 $\overline{W}=\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2$ 且 $E(W_i)=\sigma_Y^2$,所以$\frac{1}{n}\sum_{i=1}^{n}(Y_i-\overline{Y})^2\xrightarrow{p}\sigma_Y^2$。此外,$n/(n-1)\to1$,因此公式(3.7)中的第一项依概率收敛于 σ_Y^2。由于 $\overline{Y}\xrightarrow{p}\mu_Y$,$(\overline{Y}-\mu_Y)^2\xrightarrow{p}0$,因此第二项依概率收敛于 0。合并这些结果就可以得出 $s_Y^2\xrightarrow{p}\sigma_Y^2$。

第 2 部分

回归分析基础

第 4 章 一元线性回归

某一州政府对醉酒驾车的司机实行新的严厉的惩罚措施,这对高速公路交通事故死亡率会有什么影响? 一个学区减小小学的班级规模,这对该学区学生的标准化考试成绩会有什么影响? 你成功地学完了一年多的大学课程,这对你将来的收入会有多大影响?

所有这三个问题都是关于改变一个变量 X(这里 X 是指对醉酒驾车的惩罚、班级规模和受教育年数)对另一个变量 Y(这里 Y 是指交通事故死亡人数、学生考试成绩和收入水平)的未知影响。

本章介绍将一个变量 X 和另一个变量 Y 联系起来的线性回归模型。这个模型假定 X 和 Y 之间有线性关系;联系 X 和 Y 的这条直线的斜率就是 X 的单位变化对 Y 的影响。正如 Y 的均值是 Y 的总体分布的一个未知特征一样,联系 X 和 Y 的这条直线的斜率也是 X 和 Y 的总体联合分布的一个未知特征。经济计量学问题就是用这两个变量的样本数据来估计这个斜率,即估计 X 的单位变化对 Y 的影响。

本章描述了如何使用 X 和 Y 的随机样本数据对这个回归模型进行统计推断的方法。例如,使用不同学区的班级规模和考试成绩数据,我们解释了如何估计减小班级规模(比如说,每班减少 1 名学生)对考试成绩的预期影响。普通最小二乘(OLS)方法可以用来估计联系 X 和 Y 的这条直线的斜率和截距。此外,普通最小二乘估计量还可被用于检验总体斜率值的假设(例如,检验减小班级规模对考试成绩没有任何影响这一零假设)和构造斜率的置信区间。

4.1 线性回归模型

小学学区的教育主管想要决定是否雇佣更多的教师,并就此征求你的意见。如果她增加雇佣教师,她会使每个教师对应的学生数(学生—教师比)减少 2 人,因此她面临着一种权衡选择。家长希望班级的规模较小,以使他们的孩子能够得到更多的个别关注,但雇佣更多的教师意味着要花费更多的钱,而这不是那些掏钱的人愿意做的事! 所以她问你,如果她减小班级规模,对学生成绩将会有什么影响。

许多学区都采用标准化考试来测度学生的成绩，一些管理者的职位升迁或报酬多少部分地依赖于他们的学生在这些考试中的成绩表现。因此，我们可将该位主管的问题具体明确为：如果她将班级规模平均减少2名学生，那么这会对本学区标准化考试成绩产生什么样的影响？

对这个问题的精确解答需要对这里的变化进行定量的陈述。如果该教育主管以某一确定的数量改变班级的规模，那么，她期望学生的标准化考试成绩会发生怎样的变化呢？我们可用希腊字母 $\beta_{ClassSize}$ 将这种关系表示成一个数学关系，这里设置下标“*ClassSize*”的目的是把改变班级规模对学生考试成绩的影响和其他影响区别开来。因此：

$$\beta_{ClassSize}=\frac{\text{考试成绩的变化}}{\text{班级规模的变化}}=\frac{\Delta TestScore}{\Delta ClassSize} \tag{4.1}$$

其中，希腊字母 Δ（德尔塔）代表“变化”，即 $\beta_{ClassSize}$ 是由班级规模的变化所引起的考试成绩的变化，再被班级规模的变化来除。

如果你很幸运，知道 $\beta_{ClassSize}$ 的值，那么你就能够告诉教育主管，班级规模每减少1人会使整个地区的考试成绩变化 $\beta_{ClassSize}$。你还可以回答上面提到的教育主管的实际问题，即她关心的平均每个班级减少2名学生对考试成绩的影响。要回答这个问题，重新整理公式(4.1)，得到：

$$\Delta TestScore=\beta_{ClassSize}\times\Delta ClassSize \tag{4.2}$$

假如 $\beta_{ClassSize}=-0.6$，那么每班平均减少2名学生，学生考试成绩的预测变化为 $(-0.6)\times(-2)=1.2$，也就是说，如果每班减少2名学生，你就会预测到考试成绩将会平均增加1.2分。

公式(4.1)是联系考试成绩与班级规模的这一直线斜率的定义。这条直线可以写成下式：

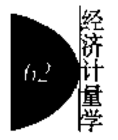
$$TestScore=\beta_0+\beta_{ClassSize}\times ClassSize \tag{4.3}$$

其中，β_0 是这条直线的截距，如前所述，$\beta_{ClassSize}$ 是直线的斜率。根据公式(4.3)，如果你知道 β_0 和 $\beta_{ClassSize}$，那么你不仅能够确定某地区与班级规模变化相联系的考试成绩的变化，你还能够预测某一给定的班级规模其平均考试成绩是多少。

当你向教育主管建议公式(4.3)时，她会告诉你这个公式有问题。她可能指出，班级规模只是影响小学教育质量多方面因素中的一个，班级规模相同的两个地区由于许多原因会有不同的考试成绩。一个地区可能拥有较好的教师，也可能使用了较好的教科书。两个具有可比较的班级规模、教师和教科书的地区还可能有非常不同的学生总体。可能一个地区有更多的移民（这样母语为英语的学生就会少）或有比较富裕的家庭。最后，她还可能指出，即使两个地区在以上所有这些方面都相同，但由于与个别学生在考试当天的表现有关的原因本质上是随机的，它们也可能会有不同的考试成绩。当然，她说的是正确的。由于所有这些原因，公式(4.3)不会对于所有的地区都精确地成立。相反，它应该被看做是平均起来在地区总体间成立的关于这种关系的一种表述。

要使这种形式的线性关系对每一个地区都成立，必须加入那些影响考试成绩的其他因素，包括每个地区独有的特征（如教师的素质、学生的背景、学生在考试那天的幸运程度等等）。一种方法是列出最重要的因素，并将它们明确地引入到公式(4.3)中（在第5章我们将介绍的理论）。但现在，我们简单地将所有这些“其他因素”并到一起，将某一特定地区的这种关系写为：

$$TestScore=\beta_0+\beta_{ClassSize}\times ClassSize+\text{其他因素} \tag{4.4}$$



这样，该地区的考试成绩受两个构成因素的影响，一个是代表学区总体中班级规模对考试成绩的平均影响，即 $\beta_0 + \beta_{ClassSize} \times$ 班级规模，另一个是用代表所有其他因素的成分即“其他因素”。

尽管这个讨论集中在考试成绩和班级规模上，但公式(4.4)中所表达的思想更具有一般意义，因此很有必要引入更一般的符号。假设有 n 个地区的样本。设 Y_i 为第 i 个地区的平均考试成绩，设 X_i 为第 i 个地区的平均班级规模，并设 u_i 为第 i 个地区影响考试成绩的其他因素，那么对每一个地区而言，即 $i = 1, \cdots, n$，公式(4.4)可被更一般地表示为：

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{4.5}$$

其中，β_0 是这条直线的截距，β_1 是斜率(公式(4.5)中的斜率用一般性符号“β_1”代替“$\beta_{ClassSize}$”，因为这个方程是按照一般性变量 X_i 来表示的)。

公式(4.5)是一元线性回归模型(linear regression model with a single regressor)，这里 Y 是因变量(dependent variable)，X 是自变量(independent variable)或回归因子(regressor)。

公式(4.5)的第一部分 $\beta_0 + \beta_1 X_i$，是总体回归线(population regression line)或总体回归函数(population regression function)。这个平均意义上的 Y 和 X 之间的关系，是针对总体而言的。因此，如果你知道 X 的值，那么根据这个总体回归线你将会预测到因变量 Y 的值是 $\beta_0 + \beta_1 X$。

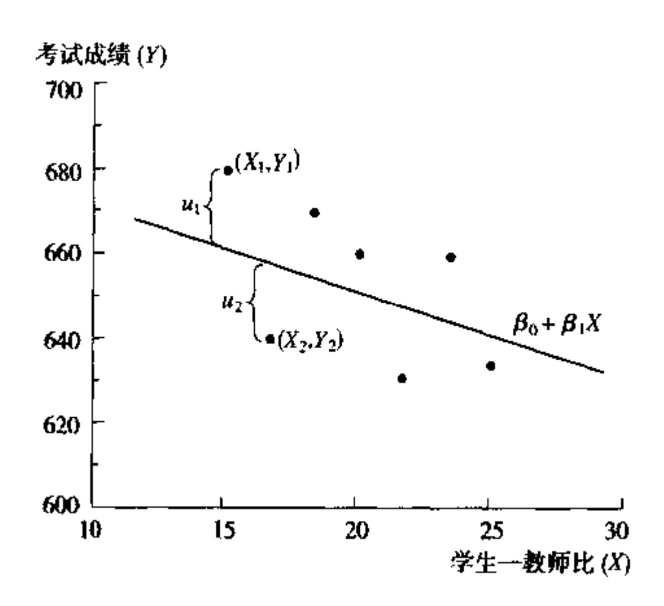
截距(intercept)β_0 和斜率(slope)β_1 是总体回归线的系数(coefficients)，也被称为总体回归线的参数(parameters)。斜率 β_1 是指与 X 的单位变化相联系的 Y 的变化。截距是指当 $X = 0$ 时，总体回归线的值，它正是总体回归线与 Y 轴的交点。在一些经济计量应用中，例如在4.7节中的应用，截距具有一定的经济意义。然而，在其他的一些应用中，截距没有现实意义。例如，当 X 代表班级规模时，严格地来讲，截距就是当班级里没有学生时考试成绩的预测值。当截距没有现实意义时，我们最好在数学上把它看做是决定总体回归线位置的系数。

公式(4.5)中的 u_i 项是误差项(error term)。误差项综合体现了造成第 i 个地区的平均考试成绩与总体回归线的预测值之间差异的所有因素。对一个具体的观测值 i 而言，误差项包含了除 X 以外的决定因变量 Y 值变化的所有其他因素。在班级规模的例子中，这些其他因素包括第 i 个地区中影响该区学生考试成绩的所有的独有特征，包括教师素质、学生的经济背景、学生的幸运程度，甚至在考试评分中的任何错误。

线性回归模型及其术语总结在重要概念4.1中给出。

图4—1描述了考试成绩(Y)和班级规模(X)之间有7个假设观测值的一元线性回归模型。总体回归线是直线 $\beta_0 + \beta_1 X$。总体回归线向下倾斜，即 $\beta_1 < 0$，这意味着具有较低的学生—教师比(较小的班级)的地区倾向于拥有较高的考试成绩。截距 β_0 作为总体回归线与 Y 轴相交的值具有数学上的意义，但正如前面所提到的，在这个例子中它没有现实意义。

由于存在决定考试成绩的其他因素，因此，图4—1中假设的观测值没有准确地落在总体回归线上。例如，地区1的 Y 值 Y_1 在总体回归线的上方。这意味着在地区1的考试成绩要比总体回归线预测得好，所以该地区的误差项 u_1 是正的。相反，Y_2 在总体回归线的下方，所以该地区考试成绩要比预测的差，$u_2 < 0$。



注：该散点图显示了7个学区的假设观测值。总体回归线是$\beta_0+\beta_1X$。从第i点到总体回归线的垂直距离是$Y_i-(\beta_0+\beta_1X_i)$，它是第i个观测值的总体误差项u_i。

图4—1　考试成绩对学生—教师比(假设数据)的散点图

重要概念4.1

一元线性回归模型的术语

一元线性回归模型是：

$$Y_i=\beta_0+\beta_1X_i+u_i$$

式中：下角标i——观测值的序号，$i=1,\cdots,n$；

Y_i——因变量，或称响应变量，或简单地称为方程的左侧变量；

X_i——自变量，或称回归因子，或简单地称为方程的右侧变量；

$\beta_0+\beta_1X$——总体回归线或总体回归函数；

β_0——总体回归线的截距；

β_1——总体回归线的斜率；

u_i——误差项。

作为一个教育主管的咨询顾问，现在回到原来的问题中：以平均每位教师减少2名学生的水平来降低学生—教师比，它对考试成绩的预期影响是什么？答案很容易：预测变化是$(-2)\times\beta_{ClassSize}$。但$\beta_{ClassSize}$的值是多少呢？

4.2　线性回归模型系数的估计

在实际情况中，比如在班级规模和考试成绩的应用例子中，总体回归线的截距β_0和斜率β_1是未知的，因此，我们必须使用数据估计总体回归线的未知斜率和截距。

这个估计问题类似于前面统计学知识复习中你曾面对过的其他问题。例如，假设你想要比较刚毕业的男女大学生的平均收入这个例子。虽然总体平均收入是未知的，但我们能够使用男女大学毕业生的随机样本估计总体均值。例如，女大学生的未知总体平均收入的自然估计量是样本中女大学毕业生的平均收入。

同样的理论可推广到线性回归模型中。我们不知道联系 X(班级规模)和 Y(考试成绩)的未知总体回归线的斜率 $\beta_{ClassSize}$ 的值,但就像可用从总体中随机抽取的样本数据来了解总体的均值一样,我们也可以利用样本数据来了解总体斜率的值即 $\beta_{ClassSize}$ 的值。

这里,我们分析的数据是由加利福尼亚州 420 个学区 1998 年从幼儿园到八年级的考试成绩和班级规模所组成的。考试成绩是整个地区五年级学生的阅读和数学的平均成绩。班级规模可以用各种方法来测度。这里用的是最广泛的一种测度,它是地区的学生数除以教师数,即整个地区的学生—教师比。附录 4.1 中更详细地描述了这些数据。

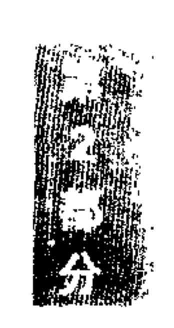
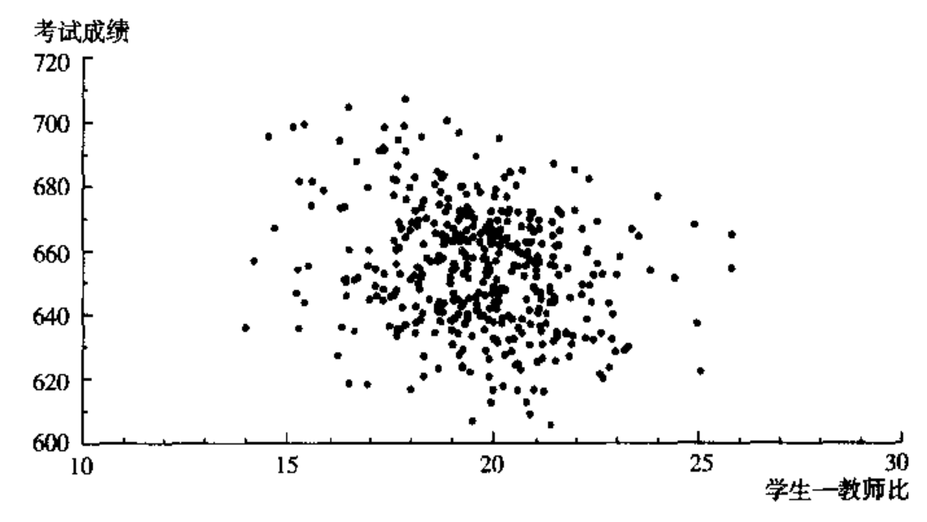


表 4—1 概括了这个样本的考试成绩和班级规模的分布。平均学生—教师比是 19.6 个学生/教师,标准差是 1.9 个学生/教师。学生—教师比分布的第 10 个百分位数是 17.3(也就是说,只有 10% 的地区的学生—教师比低于 17.3),而该地区学生—教师比的第 90 个百分位数是 21.9。

表 4—1　　1998 年加利福尼亚州 420 个 K—8 地区的五年级学生考试成绩和学生—教师比分布的汇总表

	平均数	标准差	百分位数						
			10%	25%	40%	50%(中位数)	60%	75%	90%
学生—教师比	19.6	1.9	17.3	18.6	19.3	19.7	20.1	20.9	21.9
考试成绩	654.2	19.1	630.4	640.0	649.1	654.5	659.4	666.7	679.1

关于学生考试成绩和学生—教师比的总计 420 个观测值的散点图绘制在图 4—2 中,样本相关系数是 -0.23,这表明了这两个变量之间具有弱的负相关关系。虽然在这个样本中规模较大的班级倾向于有较低的考试成绩,但是一定存在其他一些决定考试成绩的因素,正是这些因素使得观测值不能完全落在一条直线上。



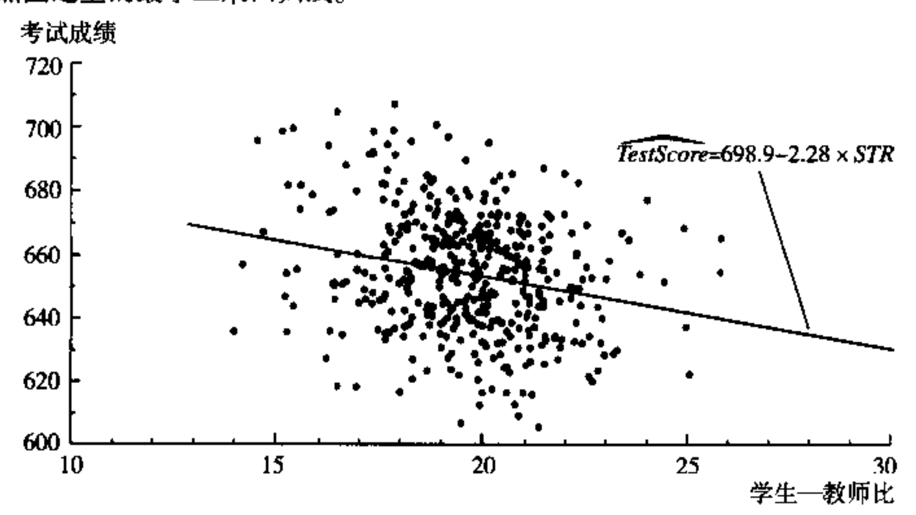
注:该数据取自于加利福尼亚州的 420 个学区。在学生—教师比和考试成绩之间存在弱的负向关系,样本相关系数是 -0.23。

图 4—2　考试成绩对学生—教师比(加利福尼亚州学区的数据)的散点图



尽管这个相关性很低,但是如果一个人能够根据这些数据画一条直线,那么基于这些数据计算出来的这条直线的斜率就可以成为 $\beta_{ClassSize}$ 的估计值。画这条直线的一种方法就是拿出铅笔和尺子并用"眼睛"画出你认为是最好的直线。虽然这种方法很容易,但非常不科学,不同的人会画出不同的估计直线。

其中,*TestScore* 是该地区的平均考试成绩,*STR* 是学生—教师比。公式(4.7)中 *TestScore* 上的符号"︿"表示这是基于普通最小二乘回归线的预测值。图4—3 绘出了加在图4—2 数据散点图之上的最小二乘回归线。



注:这条估计的回归线显示了考试成绩和学生—教师比之间的负向关系。如果班级规模减少1名学生,所估计的回归预测考试成绩会提高2.28分。

图4—3 根据加利福尼亚州数据所估计的回归线

重要概念4.2

普通最小二乘估计量、预测值和残差

斜率 β_1 和截距 β_0 的普通最小二乘估计量是:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \bar{X})(Y_i - \bar{Y})}{\sum_{i=1}^{n}(X_i - \bar{X})^2} = \frac{S_{XY}}{S_X^2} \tag{4.8}$$

$$\hat{\beta}_0 = \bar{Y} - \hat{\beta}_1\bar{X} \tag{4.9}$$

普通最小二乘估计预测值 $\hat{Y}_i$ 和残差 $\hat{u}_i$ 是:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1 X_i, i = 1, \cdots, n \tag{4.10}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i = 1, \cdots, n \tag{4.11}$$

所估计的截距($\hat{\beta}_0$)、斜率($\hat{\beta}_1$)和残差($\hat{u}_i$)是根据 X_i 和 Y_i, $i = 1, \cdots, n$ 的 n 个观测值样本计算出来的,它们是总体中未知的真实截距(β_0)、真实斜率(β_1)和真实误差项(u_i)的估计值。

-2.28 的斜率值意味着,平均来看,在一个班级中,学生—教师比每增加1个单位,整个地区的考试成绩将下降2.28分。因此,平均来看,每一个班级的学生—教师比每减少2个单位,整个地区的考试成绩将增加4.56(-2)×(-2.28),二者是有密切联系的。负的斜率表明,较大的学生—教师比(即较大的班级)与较差的考试成绩联系在一起。

现在,给定学生—教师比的值,就可以预测整个地区的考试成绩。例如,对一个学生—教师比为20的地区而言,所预测的考试成绩为698.9-2.28×20=653.3。当然,由于存在决定该地区考试成绩的其他因素,因此,这个预测不会是非常准确的。但在不考虑其他因素



的情况下，回归线的确给出了基于学生—教师比数据该地区考试成绩将会是多少的预测（普通最小二乘预测）。

这个斜率的估计值是大还是小呢？为了回答这个问题，我们回到教育主管的问题上。回想一下，她正考虑雇佣足够多的教师，以使学生—教师比减少2个单位。假设她所在地区的学生—教师比处在加利福尼亚地区的中位数上。由表4—1可知，学生—教师比的中位数是19.7，考试成绩的中位数是654.5。每个班减少2名学生，从19.7降到17.7，会使其学生—教师比从第50个百分位数移动到非常接近于第10个百分位数的位置。这是个很大的变化，她将需要雇佣许多新教师。这会对考试成绩有什么影响呢？

根据公式(4.7)，预测学生—教师比削减2个单位会使考试成绩大约增加4.6分；如果该地区的考试成绩在中位数654.5上，那么预测考试成绩将增加到659.1。这个提高是大还是小呢？根据表4—1，这个提高会使该地区从中位数移到刚好低于第60个百分位数的地方。这样，班级规模减小到使该地区接近于最小班级的10%会使考试成绩从第50个百分位数上升到第60个百分位数。根据这些估计值，大幅度削减学生—教师比(2个单位)起码是有帮助的，并且可能是值得做的，但最终要取决于她所管辖学校的财务预算情况，并且它不会是灵丹妙药。

如果这位教育主管正考虑一个更根本的变化，比如，将学生—教师比从20降到5会如何？不幸的是，公式(4.7)中的估计值对她不是很有用。这个回归是使用图4—2中的数据估计的，正如图形所示，在这些数据中最小的学生—教师比是14。这些数据不包含关于极小班级地区的学生表现的信息，所以这些数据本身并不能对这种根本性变化提供一个可靠的预测基础，即预测学生—教师比由一个普通的水平变到一个极低的水平所带来的影响。

一般兴趣框

股票的"贝塔"

现代金融学的基本思想是投资者需要有一个财务动机来承担风险。换句话说，风险投资的期望收益率[①] R 必须超过安全的或无风险投资的收益率 R_f。因此，一笔风险投资(比如持有某公司的股票)的预期超额收益率，即 $R-R_f$，应该是正的。

乍看起来，好像应该用股票的方差测度它的风险。然而，这种风险可以通过在"投资组合"中持有其他的股票来降低，也就是说，通过金融资产多样化的方式来降低大部分风险。这意味着，测度股票风险的正确方法不是用它的方差而是用它与市场的协方差。

资本资产定价模型(CAPM)将这一思想模型化。根据CAPM，一项资产的预期超额收益率与所有可获得的资产组合("市场组合")的预期超额收益率成比例。CAPM告诉我们：

$$R-R_f=\beta(R_m-R_f) \tag{4.12}$$

其中，R_m 为市场组合的期望收益率，β 是总体回归中 $R-R_f$ 对 R_m-R_f 回归的系数。实际中，无风险利率通常取短期美国政府债券的利率。根据CAPM，$\beta<1$ 的股票其风险低于市场投资组合的风险，因此应该拥有比市场投资组合低的预期超额收益率。相反，$\beta>1$ 的股票其风险高于市场投资组合的风险，因而应该要求拥有更高的预期超额收益率。

实际上，股票的"Beta"已成为投资业的一匹"载重马"，在投资公司的网站上你能够获得上百只股票的 β 估计值。这些 β 值都是用股票的实际超额收益率对综合市场指数的实际超额收益率进行普通最小二乘回归估计得到的。

表4—2给出了六只美国股票的 β 估计值。低风险的消费品生产公司具有低 β 值的股票，像Kellogg公司；高风险的高科技股票具有高的 β 值，比如微软公司。



表4—2　　六只美国股票的β值估计

公司	估计的β值
Kellogg(早餐食品公司)	0.24
Waste Management(废物处理公司)	0.38
Sprint(长途电话公司)	0.59
Walmart(折扣零售商)	0.89
Barnes and Noble(图书零售商)	1.03
Best Buy(电子设备零售商)	1.80
Microsoft(软件)	1.83

资料来源:*Yahoo.com.*

注:①一笔投资的收益率,是该笔投资的价格变化加上任何来自于该笔投资的股息或红利,被初始价格来除的百分比。例如,一只在1月1日以100美元买入的股票,在这一年中支付红利2.50美元,并在12月31日以105美元卖出,则收益率为$R=(105-100+2.50)\div100=7.5\%$。



4.2.3 为什么使用OLS估计量

使用OLS估计量$\hat{\beta}_0$和$\hat{\beta}_1$,既有实践原因,也有理论原因。因为OLS是在实践中普遍使用的方法,它已成为整个经济学、金融学(见前面的一般兴趣框)和社会科学中回归分析的最一般的通用语言。使用OLS法(或本书后面讨论的它的变形)提交报告,意味着你与其他的经济学家和统计学家"有共同语言"。实际上,OLS公式已被编程至所有的电子表格和统计软件包,使用起来非常方便。

此外,OLS估计量还具有良好的理论上的优点。例如,样本均值$\overline{Y}$是均值$E(Y)$的无偏估计量,即$E(\overline{Y})=\mu_Y$;$\overline{Y}$是μ_Y的一致性估计量;在大样本下,$\overline{Y}$的抽样分布是近似于正态分布(3.1节)。OLS估计量$\hat{\beta}_0$和$\hat{\beta}_1$也具有这些性质。在一组普通的假设下(4.3节中所陈述的),$\hat{\beta}_0$和$\hat{\beta}_1$是β_0和β_1无偏的、一致的估计量,它们的抽样分布是近似于正态的。这些结果将在4.4节中讨论。

$\overline{Y}$还有另外一个良好的理论优点,在$Y_1,\cdots,Y_n$的线性函数估计量中,它是有效的,即在所有的$Y_1,\cdots,Y_n$的加权平均估计量中,它具有最小的方差(见3.1节)。类似的结论也适用于OLS估计量,但这个结论要求另外一个假设条件,因为它超出了4.3节中的假设条件范围,所以我们将把这个讨论推迟到第4.9节中论述。

4.3 最小二乘法的假设条件

本节介绍了线性回归模型的三个假设条件和抽样方案。在这个抽样方案下,OLS为未知回归参数β_0和β_1提供了一个合适的估计量。最初,这些假设可能看起来是抽象的,不过它们确实有自然的含义。理解这些假设,对把握OLS方法什么时候能够对回归系数给出有用的估计值、什么时候给出无用的估计值,是非常必要的。



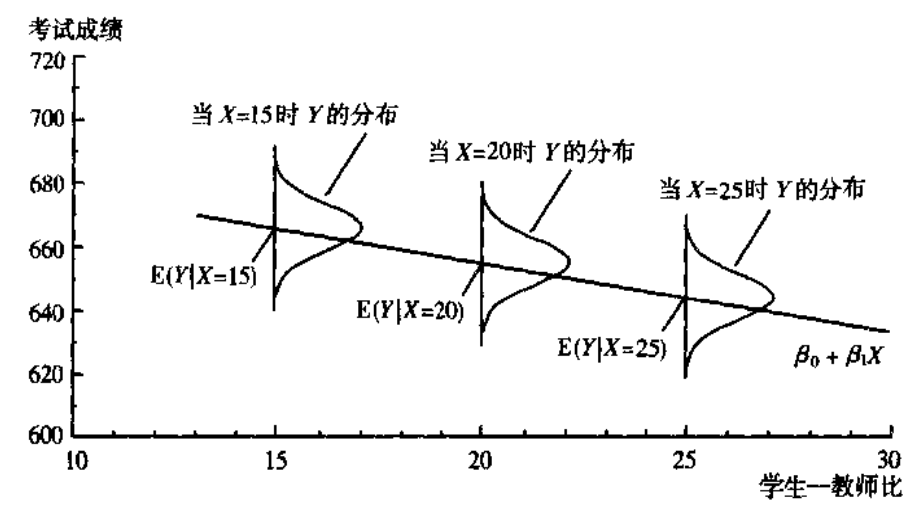
4.3.1 假设1:给定X_i,u_i的条件分布的均值为零

第一个最小二乘假设(least squares assumption)是,给定X_i条件下u_i的条件分布的

均值为零。这个假设是关于包含"其他因素"u_i 的一个正式的数学陈述,并断言这些其他因素与 X_i 无关,也就是说,给定 X_i 的值,这些其他因素分布的均值为零。

这在图4—4中进行了说明。总体回归线是平均来看总体中班级规模和考试成绩之间成立的关系,误差项 u_i 代表了导致给定地区的考试成绩不同于总体回归线预测值的其他因素。如图4—4所示,对于一个给定的班级规模的值,比如说每班20名学生,有时这些其他因素会导致比预测更好的成绩($u_i>0$),而有时还会导致比预测更差的成绩($u_i<0$),但平均来看,对总体的预测是正确的。换句话说,给定 $X_i=20$,u_i 分布的均值是零。在图4—4中,这一点被显示为 u_i 的分布集中在 $X_i=20$ 处的总体回归线的周围。更一般地讲,在 X_i 的其他值 x 处也如此,换句话说,在 $X_i=x$ 的条件下,u_i 的分布的均值为零。用数学语言表达,即 $E(u_i|X_i=x)=0$,或用更简单的符号 $E(u_i|X_i)=0$。

如图4—4所示,$E(u_i|X_i)=0$ 这一假设,与"总体回归线是给定 X_i 时 Y_i 的条件均值"(数学证明留作练习4.3)这一假设是等价的。



注:图4—4显示了班级规模为15名、20名和25名学生的学区考试成绩的条件概率。给定学生—教师比,考试成绩条件分布的均值 $E(Y|X)$ 就是总体回归线 $\beta_0+\beta_1X$。在给定 X 值之处,Y 是围绕回归线分布的,误差 $u=Y-(\beta_0+\beta_1X)$ 对所有的 X 值都具有条件零均值。

图4—4 条件概率分布和总体回归线

相关性与条件均值。回忆2.3节的内容,如果一个随机变量在给定另一个随机变量条件下的条件均值为零,那么这两个随机变量的协方差为零,进而是不相关的(公式(2.25))。因而,条件均值假设 $E(u_i|X_i)=0$ 意味着 X_i 和 u_i 是不相关的或 $\text{corr}(X_i,u_i)=0$。因为相关性是线性关联性的一种测度,所以这个含义反过来是不成立的。即使 X_i 与 u_i 不相关,给定 X_i 条件下 u_i 的条件均值可能也是非零的。不过,如果 X_i 和 u_i 是相关的,那么 $E(u_i|X_i)$ 一定是非零的。因此,就 X_i 与 u_i 之间可能的相关性来讨论条件均值假设是很方便的。如果 X_i 与 u_i 是相关的,那么条件均值假设就被违背了。

4.3.2 假设2:(X_i,Y_i),$(i=1,\cdots,n)$是独立同分布的

第二个最小二乘假设是,(X_i,Y_i),$i=1,\cdots,n$ 在观测值之间是独立同分布的(i.i.d.)。如2.5节(见重要概念2.5)中所讨论的,这一假设是关于如何抽取样本的陈述。如果使用简单随机抽样从单个大总体中抽取观测值,那么(X_i,Y_i),$i=1,\cdots,n$,就是独立同分布的。例如,设 X 是工人的年龄,Y 是他或她的收入,并且假设从工人总体中随机地抽取一个工人。

那个被随机抽取的工人会有一个确定的年龄和收入(即 X 和 Y 会取某个值)。如果从这个总体中抽取 n 个工人的样本,那么 (X_i,Y_i), $i=1,\cdots,n$,必定具有相同的分布,并且如果它们是被随机抽取的,那么它们在前后观测值之间也是独立分布的,即它们是独立同分布的。

对许多数据收集方案而言,独立同分布假设(i. i. d.)是个合理的假设。例如,从一个总体里随机选择的样本中得到的调查数据,可被典型地看做是独立同分布的。

然而,并不是所有的抽样方案得到的观测值 (X_i,Y_i) 都是独立同分布的。举一个例子,X 的值不是从总体里随机抽取的样本中得到的,而是由研究人员将其作为试验的一部分来设定的。例如,假设一位园艺家想研究不同的有机除草方法(X)对西红柿产量(Y)的影响,于是她会在种植不同西红柿的地块上施用不同的有机除草技术。如果她选择某项技术(X 的水平)用在第 i 块地上,并在所有的重复试验中将同样的技术应用于第 i 块地,那么 X_i 的值在不同样本之间是不变的。这样,X_i 是非随机的(尽管结果 Y_i 是随机的),所以该抽样方案不是独立同分布的。如果回归因子是非随机的(这将在第 15 章中进一步讨论),那么,在本章对 i. i. d. 回归因子导出的结果也同样是适用的。不过,非随机回归因子的情况是相当特殊的。例如,现代实验方案会使园艺家利用计算机化的随机数发生器把 X 的水平分配到不同的地块,因而避免了由园艺家所造成的可能偏差(她可能会对阳光充足的地块上的西红柿使用她喜欢用的除草方法)。当使用这种现代的实验方案时,X 的水平就是随机的,因此 (X_i,Y_i) 是独立同分布的。

非独立同分布抽样的另一个例子是,当观测值是从同一个观察单位在不同的时间抽取的,这时就可能产生非独立同分布的样本。例如,我们可能会有关于某公司的存货水平(Y)的数据和该公司借款利率(X)的数据,这些数据是在不同的时间从某个特定的公司中收集的。比如,这些数据可能被连续记录了 30 年,每年四次(按季度)。这是个时间序列数据的例子。时间序列数据的一个重要特征是,在时间上彼此接近的观测值之间不是独立的,而是倾向于相关的。如果现在利率水平较低,那么它们在下个季度可能也较低。这种相关性的模式,违背了独立同分布假设中"独立性"部分的内容。时间序列数据引出了一系列复杂的问题,这些问题最好是在介绍完回归分析的基本工具之后再做处理,所以我们将时间序列的进一步讨论推迟到第 4 部分。

4.3.3 假设 3:X_i 和 u_i 拥有四阶矩

第三个最小二乘假设是,X_i 和 u_i 拥有非零的有限的四阶矩($0<E(X_i^4)<+\infty$ 和 $0<E(u_i^4)<+\infty$),或者说,X_i 和 Y_i 的四阶矩是非零的和有限的。这个假设限制了抽取到极大的 X_i 或 u_i 观测值的概率。如果我们抽取到一个极大的 X_i 或 Y_i 的观测值,也就是说,X_i 或 Y_i 远在数据的正态分布范围之外,那么,这个观测值在 OLS 回归中当然会被给予很大的重视,从而导致令人误解的回归结论。

有限四阶矩的假设,在数学意义上证明了 OLS 检验统计量分布的大样本逼近是合理的。我们在第 3 章讨论样本方差一致性时曾经遇到过这个假设。具体来说,等式(3.8)表明,样本方差 s_Y^2 是总体方差 σ_Y^2 的一致估计量(即 $s_Y^2 \xrightarrow{P} \sigma_Y^2$)。如果 $Y_1,\cdots,Y_n$ 是独立同分布的,Y_i 的四阶矩是有限的,那么,在重要概念 2.6 中阐述的大数定律也适用于平均数 $\frac{1}{n}\sum_{i=1}^{n}(Y_i-\mu_Y)^2$,这是在附录 3.3 中证明 s_Y^2 是一致性估计量过程中的一个关键步骤。关于四阶矩的假设在 OLS 回归数学理论中的作用,将在 15.3 节中做进一步的讨论。

计量，即样本均值$\overline{Y}$的抽样分布的讨论。由于$\overline{Y}$是使用随机抽取的样本计算的，因此$\overline{Y}$是个随机变量，随着样本的不同而取不同的值，这些不同值的概率体现在它的抽样分布中。尽管当样本容量很小时，$\overline{Y}$的抽样分布可能是复杂的，但给出对于所有的n都成立的确定性的分布陈述却是可能的，尤其是抽样分布的均值是μ_Y，即$E(\overline{Y})=\mu_Y$，所以$\overline{Y}$是μ_Y的无偏估计量。如果n很大，那么对于抽样分布能够总结的结论更多，尤其重要的是，中心极限定理（见2.6节）表明了这个分布是渐近正态的。

$\hat{\beta}_0$和$\hat{\beta}_1$的抽样分布　以上这些理论，仍然可以应用到总体回归线未知截距β_0和未知斜率β_1的OLS估计量$\hat{\beta}_0$和$\hat{\beta}_1$的分布上。由于OLS估计量是使用随机样本计算的，因此$\hat{\beta}_0$和$\hat{\beta}_1$是随机变量，随着样本的不同而取不同的值，这些不同值的概率体现在它们的抽样分布中。



尽管当样本容量很小时，$\hat{\beta}_0$和$\hat{\beta}_1$的抽样分布可能是复杂的，但给出对于所有的n都成立的确定性的分布陈述却是可能的。特别是$\hat{\beta}_0$和$\hat{\beta}_1$的抽样分布的均值是β_0和β_1。换句话说，在重要概念4.3中的最小二乘假设下：

$$E(\hat{\beta}_0)=\beta_0,E(\hat{\beta}_1)=\beta_1 \tag{4.13}$$

即$\hat{\beta}_0$和$\hat{\beta}_1$是β_0和β_1的无偏估计量。关于$\hat{\beta}_1$是无偏的这一结论的证明在附录4.3中给出，而关于$\hat{\beta}_0$是无偏的这一结论的证明留作练习4.4。

如果样本足够大，那么根据中心极限定理，$\hat{\beta}_0$和$\hat{\beta}_1$的抽样分布可很好地被二元正态分布逼近（见2.4节）。这意味着在大样本条件下，$\hat{\beta}_0$和$\hat{\beta}_1$的边缘分布是正态的。

这个论证需要借助于中心极限定理。从技术上说，中心极限定理主要涉及平均值（比如$\overline{Y}$）的分布。如果你研究一下公式(4.8)中$\hat{\beta}_1$的分子，你就会发现它也是一种平均数——不是像$\overline{Y}$那样的简单平均数，而是一个乘积$(Y_i-\overline{Y})(X_i-\overline{X})$的平均数。如同附录4.3中进一步讨论的，中心极限定理也适用于这个平均数，像简单平均数$\overline{Y}$一样，在大样本条件下它也服从正态分布。

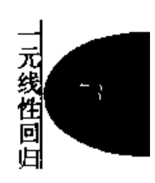
关于大样本条件下OLS估计量分布的正态近似总结在重要概念4.4中给出（附录4.3中概要地给出了这些公式的推导）。一个实践中的重要问题是，n取多大时这些近似才是可靠的。在2.6节中，对于$\overline{Y}$的抽样分布怎样才能很好地被正态分布所逼近这一问题，我们建议$n=100$是足够大的，有时较小的n也能满足要求。这个准则可继续应用到回归分析中出现的更加复杂的平均数上。实际上在所有现代经济计量应用中，一般总是有$n>100$，所以我们将把OLS估计量分布的正态近似看做是可靠的，除非有很好的理由认为它不是正态分布。

重要概念4.4

$\hat{\beta}_0$和$\hat{\beta}_1$的大样本分布

如果重要概念4.3中的最小二乘假设成立，那么在大样本条件下，$\hat{\beta}_0$和$\hat{\beta}_1$服从联合正态抽样分布。$\hat{\beta}_1$的大样本正态分布是$N(\beta_1,\sigma^2_{\hat{\beta}_1})$，其中这个分布的方差$\sigma^2_{\hat{\beta}_1}$为：

$$\sigma^2_{\hat{\beta}_1}=\frac{1}{n}\frac{\mathrm{var}[(X_i-\mu_X)u_i]}{[\mathrm{var}(X_i)]^2} \tag{4.14}$$

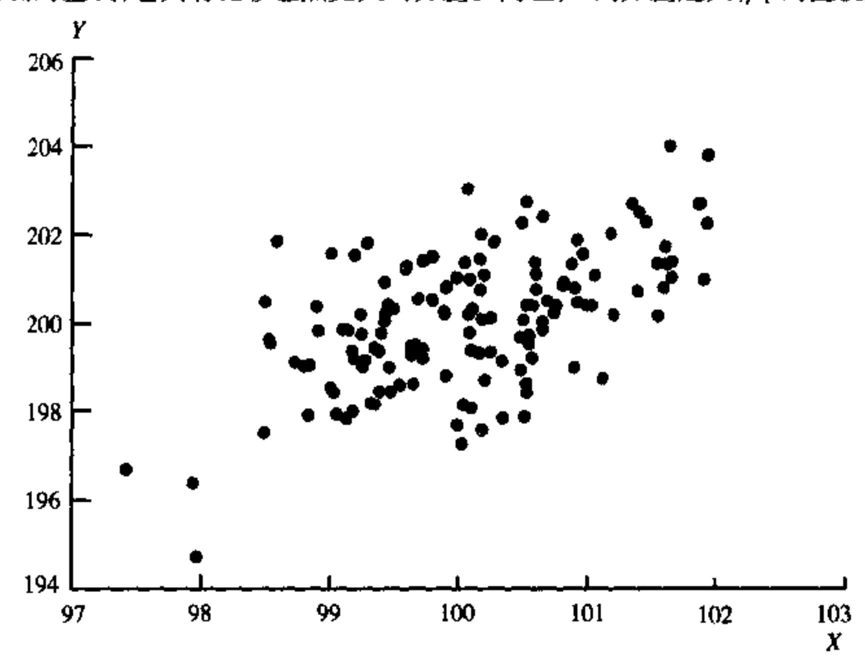


$\hat{\beta}_0$的大样本正态分布是$N(\beta_0,\sigma^2_{\hat{\beta}_0})$，其中：

$$\sigma^2_{\hat{\beta}_0}=\frac{1}{n}\frac{\mathrm{var}(H_iu_i)}{[E(H_i^2)]^2},\text{这里 } H_i=1-\left(\frac{\mu_X}{E(X_i^2)}\right)X_i \tag{4.15}$$

重要概念 4.4 中总结的结果表明，OLS 估计量是一致的，也就是说，当样本很大时，$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 将会以很高的概率逼近于真实总体系数 β_0 和 β_1。这是因为随着 n 的增大，估计量的方差 $\sigma^2_{\hat{\beta}_0}$ 和 $\sigma^2_{\hat{\beta}_1}$ 减少到零（n 出现在方差表达式的分母中）。所以当 n 很大时，OLS 估计量的分布将紧紧地集中在它们的均值 β_0 和 β_1 的周围。

重要概念 4.4 中给出的大样本分布的另一个意义是，一般来说 X_i 的方差越大，$\hat{\beta}_1$ 的方差 $\sigma^2_{\hat{\beta}_1}$ 就越小。从数学上看，因为等式（4.14）中 $\hat{\beta}_1$ 的方差与 X_i 的方差平方成反比，所以，$\mathrm{var}(X_i)$ 越大，等式（4.14）中的分母就越大，$\sigma^2_{\hat{\beta}_1}$ 就越小。为了更好地理解为什么会是这样的原因，请看图 4—5，这个图绘出了关于 X 和 Y 的 150 个人工数据点。用彩色点表示的数据点是最接近 $\bar{X}$ 的 75 个观测值。假如让你尽可能精确地画出一条经过彩色点或黑色点的直线，你是选择经过彩色点的直线还是选择经过黑色点的直线？通过黑色点会比较容易地画出一条精确的直线，它具有比彩色点更大的方差。同理，X 的方差越大，$\hat{\beta}_1$ 的值就越精确。



注：彩色点代表具有小方差的一组 X_i 值，黑色点代表具有大方差的一组 X_i 值。黑色点会比彩色点更精确地估计出回归线。

图 4—5　$\hat{\beta}_1$ 的方差和 X 的方差

$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的抽样分布正态近似是一个非常有力的工具。有了这个近似，我们就能够设计出只使用样本数据就能够对回归系数的总体真值进行推断的方法。

4.5　检验单个回归系数的假设



你的一个客户，也即前例中的那位教育主管，给你打电话咨询一个问题。在她的办公室里正有一位愤怒的纳税人，他声称削减班级规模对考试成绩没有帮助，因此进一步地减小班级规模是在浪费资金。这位纳税人声称班级规模对考试成绩没有影响。

这位纳税人的主张能用回归分析的语言重新表述。由于班级规模的单位变化对考试成绩的影响是 $\beta_{ClassSize}$，因此这位纳税人声称的结论意味着总体回归线是平坦的，也就是说，总

$$\hat{\sigma}^2_{\hat{\beta}_1}=\frac{1}{n}\times\frac{\frac{1}{n-2}\sum_{i=1}^{n}(X_i-\bar{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2\right]^2} \tag{4.19}$$

公式(4.19)中的方差估计量在附录4.4中做了讨论。尽管$\hat{\sigma}^2_{\hat{\beta}_1}$的表达式很复杂,但因为在应用中用回归软件计算标准误,所以在实际中它是比较容易计算的。

第二步是计算t统计量。

$$t=\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)} \tag{4.20}$$

第三步是计算p值,即当假定零假设为真时,观察到的$\hat{\beta}_1$值至少和实际计算的估计值($\hat{\beta}_1^{act}$)一样不同于$\beta_{1,0}$的概率。数学上表述为:

$$p\text{-值}=\Pr_{H_0}[|\hat{\beta}_1-\beta_{1,0}|>\hat{\beta}_1^{act}-\beta_{1,0}]$$

$$=\Pr_{H_0}\left[\left|\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)}\right|>\left|\frac{\hat{\beta}_1^{act}-\beta_{1,0}}{SE(\hat{\beta}_1)}\right|\right]=\Pr_{H_0}(|t|>|t^{act}|) \tag{4.21}$$

其中,$\Pr_{H_0}$表示在零假设下所计算的概率,第二个等式通过除以$SE(\hat{\beta}_1)$得到,而t^{act}为实际所计算的t统计量的值。因为在大样本条件下,$\hat{\beta}_1$服从渐近正态分布,所以在零假设下,t统计量是个渐近服从标准正态分布的随机变量。因此,在大样本条件下有:

$$p\text{值}=\Pr(|Z|>|t^{act}|)=2\Phi(-|t^{act}|) \tag{4.22}$$

一个很小的p值,比如说小于5%,提供了反对零假设的证据,其意义是说,如果实际上零假设是真的,则从各个纯随机抽取的样本中获得的$\hat{\beta}_1$值的机会小于5%。如果是这样的话,零假设则在5%的显著性水平下被拒绝。

另一种方法是,通过简单地比较t统计量的值和双边检验的临界值±1.96,在5%的显著性水平下检验零假设。如果$|t^{act}|>1.96$,那么就在5%的显著性水平下拒绝零假设。

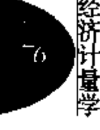
这几个步骤被总结在重要概念4.6中。

应用于考试成绩这个例子。使用图4—2中的420个观测值所估计的、并在公式(4.7)中所报告的斜率系数的OLS估计量是-2.28。它的标准误是0.52,即$SE(\hat{\beta}_1)=0.52$。因而,为了检验零假设$\beta_{ClassSize}=0$,我们使用公式(4.20)构造t统计量。因此,$t^{act}=(-2.28-0)\div0.52=-4.38$。

这个t统计量超过了1%的双边临界值2.58,所以在1%的显著水平下拒绝零假设,支持备择假设。另一方面,我们可以计算与$t=-4.38$相联系的p值。这个概率是标准正态分布尾部的面积,如图4—6所示。这个概率值非常小,约为0.00001或0.001%。也就是说,如果零假设$\beta_{ClassSize}=0$是真的,那么得到一个和我们实际得到的值一样远离零假设的$\hat{\beta}_1$值的概率极其小,小于0.001%。由于这个事件是如此地不可能,因此有理由得出结论:零假设是假的。

4.5.2 关于β_1的单边假设

到目前为止,我们讨论的重点一直集中在零假设为$\beta_1=\beta_{1,0}$和备择假设为$\beta_1\neq\beta_{1,0}$的假设检验。这是一种双边假设检验,因为在备择假设下,β_1要么可能大于$\beta_{1,0}$,要么可能小于$\beta_{1,0}$。但是,有时用单边假设检验是比较合适的。例如,在学生—教师比和考试成绩的问题



中，许多人认为规模较小的班级能够给学生提供一个更好的学习环境。在此假设下，β_1 是负的，表明规模较小的班级会导致较高的成绩。因此，检验零假设 $\beta_1=0$（没有影响）和单边备择假设 $\beta_1<0$，应该是有意义的。

对单边检验而言，零假设和单边备择假设是：

$$H_0:\beta_1=\beta_{1,0}, H_1:\beta_1<\beta_{1,0}（单边备择假设） \tag{4.23}$$

其中，$\beta_{1,0}$是在零假设下 β_1 的值（在学生—教师比的例子中为0），而备择假设是 β_1 小于 $\beta_{1,0}$。如果备择假设是 β_1 大于 $\beta_{1,0}$，那么表达式(4.23)中不等式要改变方向。



重要概念 4.6

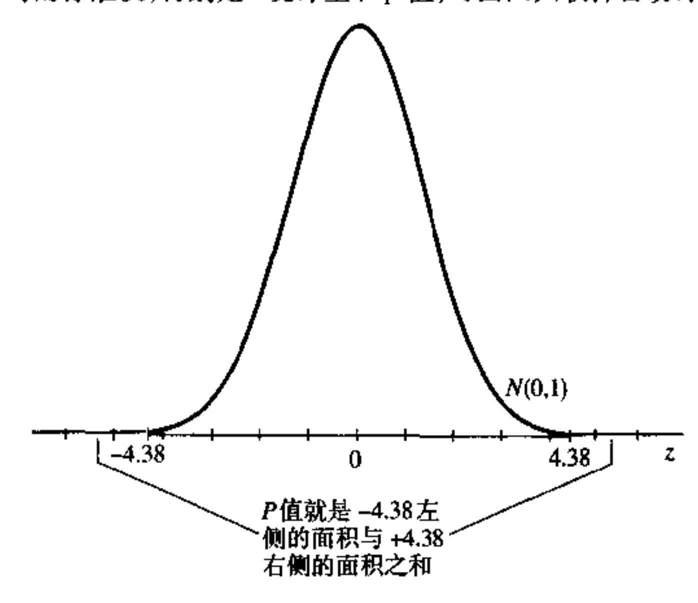
零假设 $\beta_1=\beta_{1,0}$和备择假设 $\beta_1\neq\beta_{1,0}$的检验

1. 计算 $\hat{\beta}_1$ 的标准误 $SE(\hat{\beta}_1)$（公式(4.18)）。

2. 计算 t 统计量（公式(4.20)）。

3. 计算 p 值（公式(4.22)）。如果 p 值小于 0.05 或者 $|t^{act}|>1.96$，那么在 5% 的显著水平下拒绝零假设。

检验 $\beta_1=0$ 时的标准误，特别是 t 统计量和 p 值，可由回归软件自动计算得出。



注：双边检验的 p 值是 $|Z|>|t^{act}|$ 的概率，其中 Z 是个标准正态随机变量，而 t^{act} 是根据样本所计算的 t 统计量的值。当 $t^{act}=-4.38$ 时，p 值仅为 0.00001。

图 4—6 计算当 $t^{act}=-4.38$ 时双边检验的 p 值

由于零假设对单边和双边假设检验都是相同的，因此，t 统计量的构造也是相同的。单边和双边假设检验之间的惟一区别在于如何解释 t 统计量。对表达式(4.23)中的单边备择假设而言，在单边备择假设下，t 统计量很大的负值（不是较大的正值）将拒绝零假设：如果 $t^{act}<-1.645$，那么在5%的显著水平下拒绝零假设，而在 $|t^{act}|>1.96$ 时则不能拒绝零假设。

由累积标准正态分布可获得单边检验的 p 值，即：

$$p\text{ 值}=\Pr(Z<t^{act})=\Phi(t^{act}) \quad (p\text{ 值，单边左尾检验}) \tag{4.24}$$

如果备择假设是 $\beta_1>\beta_{1,0}$，那么表达式(4.23)和表达式(4.24)中的不等式要改变方向，这样 p 值是右尾概率，$\Pr(Z>t^{act})$。



单边检验在什么情况下使用？在实际应用中，只有当具备清晰的理由认为 β_1 确实是在

零假设值$\beta_{1,0}$的一侧时，才应该使用单边备择假设。这个理由可能源自于经济理论、先前的经验证据或二者的结合，然而，即使相应的备择假设在最初好像是单边的，但实际上它可能并不一定如此。比如对于一种新研制的药品来说，由于先前可能没被意识到的某种副作用，在实际临床试验中该药品被证实是有害的。在班级规模的例子中，我们想起这样一个关于大学毕业生的笑话，据传一所大学成功的秘密在于招收天才的学生，并确保教职员工远离他们，尽可能少伤害他们。在实际中，这种模棱两可性经常使得经济计量学家采用双边检验。

在考试成绩案例中的应用。检验班级规模对考试成绩没有影响（即表达式（4.23）中的$\beta_{1,0}=0$）这一假设的t统计量为$t^{act}=-4.38$。它小于-2.33（1%显著性水平下单边检验的临界值），所以对应于单边备择假设的零假设在1%的显著性水平下被拒绝。事实上，p值小于0.0006%。根据这些数据，你就可以在1%的显著水平下拒绝那位愤怒的纳税人的断言，即拒绝认为斜率的负估计值纯粹是由于随机抽样变化所引起的。

4.5.3 检验关于截距项β_0的假设

前面的讨论一直集中在检验关于斜率β_1的假设上，然而，有时也涉及截距β_0假设的检验。关于截距的零假设和双边备择假设是：

$$H_0:\beta_0=\beta_{0,0},H_1:\beta_0\neq\beta_{0,0}\text{（双边备择假设）}\tag{4.25}$$

前面介绍的检验零假设的三个步骤也同样适用于β_0的检验（见重要概念4.6）（$\hat{\beta}_0$的标准误表达式在附录4.4中给出）。如果备择假设是单边的，那么这种方法要进行修正，就像上小节中所讨论的有关斜率的假设检验一样。

如果你心中有个特定的零假设（比如那位愤怒的纳税人），那么假设检验是很有用的。在统计证据的基础上接受或拒绝这个零假设，为处理利用样本信息了解总体特征过程中产生的不确定性这一问题提供了一个有力的工具。然而，很多的时候关于回归系数的单一假设没有多大市场，相反，人们更想要知道基于数据之上系数取值的某一范围。这样就要求我们构造置信区间。

4.6 回归系数的置信区间

由于斜率β_1的任何统计估计值必然会含有抽样不确定性，因此根据样本数据我们不可能精确地确定β_1的真实值。但是我们却可以使用普通最小二乘估计量及其标准误来构造斜率β_1或截距β_0的置信区间。

β_1的置信区间。回想一下，β_1的95%的置信区间有两个等价的定义。第一，它是使用双边假设检验在5%的显著性水平下不能被拒绝的值的集合。第二，它是一个以95%的概率包含β_1的真实值的区间。也就是说，在被抽取的95%的可能样本中，置信区间将包含β_1的真实值。因为这个置信区间包含了所有样本中95%的样本的真实值，所以它被称为具有95%的置信水平。



这两个定义等价的原因如下。根据定义，5%的显著性水平下的假设检验在所有可能样本中只有5%的机会拒绝β_1的真实值，也就是说，在95%的所有可能样本中，β_1的真实值将不会被拒绝。因为95%的置信区间（如第一个定义所定义的）是在5%的显著性水平下不能被拒绝的β_1的所有值的集合，所以可以进一步推论，在所有可能的样本中有95%的机会使β_1的真实值被包含在置信区间里。

和总体均值置信区间的估计一样(见3.3节),原则上,能够在5%的显著水平下通过使用t统计量检验β_1的所有可能值(即对所有的$\beta_{1,0}$值,检验零假设$\beta_1=\beta_{1,0}$)计算95%的置信区间。那么,95%的置信区间就是不能被拒绝的所有β_1值的集合,但对所有的β_1值构造t统计量将会很费劲。

构造置信区间的一种简单方法,是应该注意到只要$\beta_{1,0}$落在范围$\hat{\beta}_1 \pm SE(\hat{\beta}_1)$之外,$t$统计量就会拒绝假设的$\beta_{1,0}$值。也就是说,$\beta_1$的95%的置信区间就是区间$(\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1))$。这一方法与总体均值置信区间的估计方法是类似的。

β_1的置信区间的构造方法总结在重要概念4.7中给出。

β_0的置信区间。β_0的95%的置信区间的构造方法与重要概念4.7中介绍的方法一样,不过需要用$\hat{\beta}_0$和$SE(\hat{\beta}_0)$分别替换$\hat{\beta}_1$和$SE(\hat{\beta}_1)$。

在考试成绩案例中的应用。公式(4.7)所给出的考试成绩对学生—教师比的OLS回归结果是:$\hat{\beta}_0=698.7$,$\hat{\beta}_1=-2.28$。这些估计值的标准误为$SE(\hat{\beta}_0)=10.4$,$SE(\hat{\beta}_1)=0.52$。

由于标准误的重要性,今后在给出OLS回归线时,我们会把标准误包括在估计系数下面的括号内:

$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR \tag{4.26}$$

β_1的95%的双边置信区间是$\{-2.28 \pm 1.96 \times 0.52\}$,即$-3.30 \leqslant \beta_1 \leqslant -1.26$。$\beta_1=0$的值并不包含在这个置信区间内,因此(如我们在4.5节中已知道的),$\beta_1=0$的假设在5%的显著性水平下被拒绝。



重要概念4.7

β_1的置信区间

β_1的95%的双边置信区间是以95%的概率包含β_1真实值的区间,也就是说,它在95%的所有可能随机抽取的样本中包含β_1的真实值。等价地,它是不能被5%的双边假设检验拒绝的所有β_1的值的集合。当样本容量很大时,它被构造为:

$$\beta_1 \text{的95\%的置信区间} = (\hat{\beta}_1 - 1.96SE(\hat{\beta}_1), \hat{\beta}_1 + 1.96SE(\hat{\beta}_1)) \tag{4.27}$$

改变X所导致的预测效应的置信区间。β_1的95%的置信区间能够用于构造X的普通变化所导致的预测效应的95%的置信区间。

考虑按一个给定的量Δx改变X。与X的这个变化相联系的Y的预测变化值为$\beta_1\Delta x$。虽然总体斜率β_1是未知的,但是因为我们能够构造β_1的置信区间,所以,我们就能够构造预测效应$\beta_1\Delta x$的置信区间。由于β_1的95%的置信区间一端的值是$\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)$,因此,使用$\beta_1$的该估计值,变化$\Delta x$的预测效应为$(\hat{\beta}_1 - 1.96SE(\hat{\beta}_1)) \times \Delta x$。置信区间另一端的值是$\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)$,使用该估计值,变化后的预测效应为$[(\hat{\beta}_1 + 1.96SE(\hat{\beta}_1)] \times \Delta x$。因此,以一定量$\Delta x$改变$x$所产生的效应的95%的置信区间能被表示为:

$$\beta_1\Delta x \text{的95\%的置信区间} = (\hat{\beta}_1\Delta x - 1.96SE(\hat{\beta}_1)\Delta x, \hat{\beta}_1\Delta x + 1.96SE(\hat{\beta}_1)\Delta x) \tag{4.28}$$



例如,我们假想的那位教育主管正考虑减少学生—教师比,即平均每名教师减少2名学生。由于β_1的95%的置信区间是$(-3.30, -1.26)$,因此,把学生—教师比减少2个单位所产生的效应,最大可能值为$(-3.30)\times(-2)=6.60$,最小可能值为$(-1.26)\times(-2)=$

验这两个总体均值相同的零假设和对应的备择假设,等同于检验零假设$\beta_1=0$和双边备择假设$\beta_1\neq0$。这个假设能够用4.5节中所列出的程序进行检验。具体来说,当OLS的t统计量的值$t=\hat{\beta}_1/SE(\hat{\beta}_1)$的绝对值超过1.96时,对应于双边检验备择假设的零假设在5%的显著性水平下被拒绝。同理,如4.6节中所描述的,β_1的95%的置信区间即$\hat{\beta}_1\pm1.96SE(\hat{\beta}_1)$,为两个总体均值的差提供了一个95%的置信区间。

在考试成绩案例中的应用　作为一个例子,利用图4—2中的420个观测值,使用OLS法估计考试成绩对定义在(4.29)中的学生—教师比二元变量D进行回归,结果为:

$$\widehat{TestScore} = \underset{(1.3)}{650.0} - \underset{(1.8)}{7.4}D \tag{4.33}$$

其中,系数β_0和β_1的OLS估计值的标准误在OLS估计值下面的括号内给出。因此,学生—教师比大于或等于20(即$D=0$)的子样本的平均考试成绩是650.0,而学生—教师比小于20(即$D=1$)的子样本的平均考试成绩是$650.0+7.4=657.4$。因而,这两组样本平均考试成绩之差是7.4。这就是学生—教师比二元变量D的系数β_1的OLS估计值。

这两个组的总体平均考试成绩之差统计上在5%的显著性水平下是否显著地异于0?为了得出结果,构造关于β_1的t统计量:$t=7.4/1.8=4.04$。它在绝对值上大于1.96,所以在5%的显著性水平下,我们拒绝“学生—教师比高的地区与学生—教师比低的地区总体平均考试成绩相同”这一零假设。

OLS估计量及其标准误可用来构造均值真实差异的95%的置信区间。这个置信区间是:$7.4\pm1.96\times1.8=(3.9,10.9)$。这个置信区间不包括$\beta_1=0$,所以(根据上一段我们所知道的)我们在5%的显著性水平下拒绝$\beta_1=0$这一假设。

4.8　R^2 和回归的标准误

R^2和回归的标准误是评价OLS回归线对数据拟合程度好坏的两个测度指标。R^2的值在0和1之间变化,它测度了Y_i的方差中由X_i的变化所解释的部分。回归标准误则直观地测度了Y_i距离它的预测值有多远。

4.8.1　R^2

回归的R^2(regression R^2)是指在Y_i的样本方差中由X_i所解释(或预测)的部分。预测值和残差的定义(见重要概念4.2)允许我们将因变量Y_i表示为预测值$\hat{Y}_i$加上残差$\hat{u}_i$之和:

$$Y_i=\hat{Y}_i+\hat{u}_i \tag{4.34}$$

在这个符号表示中,R^2是$\hat{Y}_i$的样本方差与Y_i的样本方差之比。

在数学上,R^2可被表示为被解释的平方和与总平方和之比。被解释的平方和(explained sum of squares)或ESS,等于Y_i的预测值$\hat{Y}_i$与其均值的离差平方和;总平方和(total sum of squares)或TSS,等于Y_i与其均值的离差平方和,即:

$$ESS=\sum_{i=1}^{n}(\hat{Y}_i-\bar{Y})^2 \tag{4.35}$$

$$TSS=\sum_{i=1}^{n}(Y_i-\bar{Y})^2 \tag{4.36}$$





其中，公式(4.35)使用了 $\bar{Y}$ 等于样本均值的 OLS 预测值这一事实(证明见附录4.3)。

R^2 是被解释的平方和与总平方和之比，即：

$$R^2 = \frac{ESS}{TSS} \tag{4.37}$$

换一种表示方法，R^2 也可表示为 Y_i 的方差中不能由 X_i 所解释的那部分方差的比重。残差平方和(sum of squares residuals)或 *SSR*，等于 OLS 残差平方和，即：

$$SSR = \sum_{i=1}^{n} \hat{u}_i^2 \tag{4.38}$$

在附录4.3中证明了 $TSS = ESS + SSR$。因此，R^2 也能被表示为1减去残差平方和与总平方和之比，即：

$$R^2 = 1 - \frac{SSR}{RSS} \tag{4.39}$$

最后，Y 对单个回归因子 X 回归的 R^2 是 Y 和 X 之间相关系数的平方。

R^2 的取值范围在0和1之间。如果 $\hat{\beta}_1 = 0$，那么 X_i 没有解释 Y_i 的方差，而基于回归的 Y_i 的预测值恰好是 Y_i 的样本均值。在这种情况下，被解释的平方和为0，残差平方和等于总平方和，因而，R^2 为0。相反，如果 X_i 解释了 Y_i 的所有变化，那么对于所有的 i，有 $Y_i = \hat{Y}_i$，每个残差都为0(即 $\hat{u}_i = 0$)，这样，就有 $ESS = TSS$ 和 $R^2 = 1$。一般地说，R^2 不会取0或1这两个极值，而是取二者之间的某个值。R^2 接近于1，表明回归因子预测 Y_i 的效果很好，而 R^2 接近于0，表明回归因子预测 Y_i 的效果很差。

4.8.2 回归的标准误

回归的标准误(standard error of the regression)或 *SER* 是回归误差 u_i 的标准差的估计量。由于回归误差 $u_1, \cdots, u_n$ 是观察不到的，因此，用对应的样本残差即 OLS 残差 $\hat{u}_1, \cdots, \hat{u}_n$ 来计算 *SER*。*SER* 的计算公式如下：

$$SER = s_{\hat{u}}，其中，s_{\hat{u}}^2 = \frac{1}{n-2}\sum_{i=1}^{n} \hat{u}_i^2 = \frac{SSR}{n-2} \tag{4.40}$$

这里，$s_{\hat{u}}^2$ 的表达式用到了 OLS 残差的样本均值为零这个事实(证明见附录4.3)。

除了公式(3.7)中的 $Y_i - \bar{Y}$ 被 $\hat{u}_i$ 所代替，公式(3.7)中的除数是 $n-1$ 而这里的除数是 $n-2$ 以外，公式(4.40)中 *SER* 的表达式和3.2节里公式(3.7)中所给出的 Y 的样本标准差的表达式是相同的。这里除数用 $n-2$(代替 n)的原因与公式(3.7)中的除数用 $n-1$ 的原因一样：它修正了由估计两个回归系数所引起的微小的向下偏差，这被称为"自由度"修正；由于要估计两个系数(β_0 和 β_1)，因此损失了数据的两个自由度，这样这个因子中的除数就是 $n-2$(其背后的数学意义在15.4节中讨论)。当 n 很大时，除以 n、$n-1$ 或 $n-2$ 的差别是可以忽略的。

4.9 异方差性和同方差性

关于以 X_i 为条件的 u_i 的分布，我们只假设它具有零均值(第一个最小二乘假设)。此外，如果这个条件分布的方差不依赖于 X_i，那么就称误差项是同方差的。本节讨论了同方差性及其理论含义，在误差是同方差条件下 OLS 估计量的标准误的简化表达式，以及在实际中使用这些简化表达式所面临的风险。

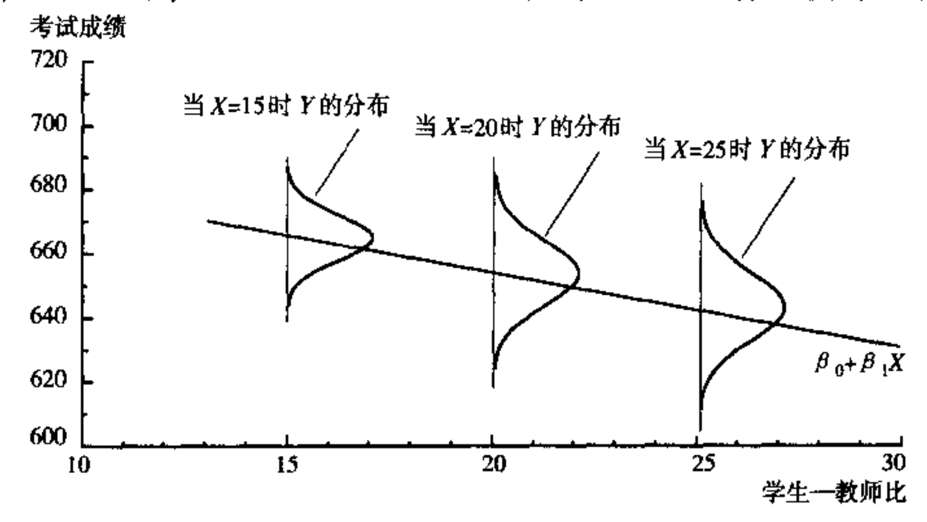


4.9.1　什么是异方差和同方差

异方差和同方差的定义。如果给定 X_i，对于任何的 $i=1,n,u_i$ 的条件分布的方差为常数，尤其是它不依赖于 X_i，那么就称误差项 u_i 是同方差的(homoskedastic)；否则，误差项就是异方差的(heteroskedastic)。

作为一个例子，我们回到图4—4。图4—4 给出了误差项 u_i 在不同 x 值条件下的分布。由于这个分布适用于特别指定的 x 值，因此，这就是在给定 $X_i=x$ 条件下 u_i 的条件分布。正如该图所描述的，所有这些条件分布都具有相同的离散度，更精确地说，对于不同的 x 值，这些分布的方差都是相同的。也就是说，在图4—4 中，给定 $X_i=x$，u_i 的条件方差不依赖于 x 的变化而变化，所以图4—4 中所说明的误差是同方差的。

相反，图4—7 列示了 u_i 的条件分布随着 x 的增加而向外扩散的一种情况。对于较小的 x 值，这个分布是密集的，但对于较大的 x 值，它具有更大的离散度。因此在图4—7 中，给定 $X_i=x$ 条件下，u_i 的方差随着 x 的增加而增大，这样图4—7 中的误差就是异方差的。



注：像图4—4 一样，这个图显示了三个不同规模班级的考试成绩的条件分布。和图4—4 不一样的是，对较大的班级规模而言，这些分布更加向外扩展(具有较大的方差)。由于给定 X，u 的分布的方差 $var(u/X)$ 依赖于 X，因此 u 是异方差的。

图4—7　异方差的一个例子

异方差和同方差的定义总结在重要概念4.8 中给出。

重要概念4.8

异方差和同方差

对于任何的 $i=1,\cdots,n$，如果给定 X_i 条件下 u_i 的条件分布的方差 $var(u_i|X_i=x)$ 都是常数，尤其是它不依赖于 x 的变化而变化，那么就称误差项 u_i 是同方差的；否则，误差项就是异方差的。

例子。这两个术语的构词是很长的，而且这个定义可能很抽象。为了用一个例子阐明这两个术语，我们离开一下学生—教师比与考试成绩的问题，而回到3.5 节中所考虑的男女大学毕业生收入的例子中。设 $MALE_i$ 是个二元变量，对男大学生而言，它等于1，而对女大学生而言，它等于0。对 $i=1,\cdots,n$，联系某人的收入及其性别的二元变量回归模型为：

$$Earnings_i=\beta_0+\beta_1 MALE_i+u_i \tag{4.41}$$



误差项是同方差的，那么存在一个专门的公式用于计算 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的标准误，这些公式在附录 4.4 中给出。在 X 是二元变量这一特殊情况下，在同方差下，$\hat{\beta}_1$ 的方差估计量（即在同方差下 $\hat{\beta}_1$ 的标准误的平方）就是在 3.4 节的脚注①中讨论的所谓的均值差异的混合方差公式（pooled variance formula）。

由于这些可供选择的公式都是在误差项是同方差的特定情况下推导得出的，并不适用于误差项是异方差的情况，所以它们被称为 OLS 估计量的方差和标准误的"仅适用于同方差的"公式。正像这个名称所暗示的，如果误差项是异方差的，那么这个仅适用于同方差的标准误（homoskedasticity－only stardard errors）就是不适当的。特别是说，如果误差项是异方差的，那么使用仅适用于同方差的标准误计算的 t 统计量不服从标准正态分布，即使在大样本条件下也是如此。事实上，用于这个仅适用于同方差的 t 统计量的正确的临界值取决于这个异方差的确切性质，所以这些临界值不能被列表给出。同理，如果误差项是异方差的，但置信区间被构造为 ±1.96 倍的仅适用于同方差的标准误，那么一般而言，这个区间包含系数真实值的概率不是 95%，即使在大样本条件下也是如此。



相反，由于同方差是异方差的一个特例，因此，不论误差项是异方差的还是同方差的，公式(4.19)和公式(4.59)中给出的 $\hat{\beta}_1$ 和 $\hat{\beta}_0$ 的方差估计量 $\hat{\sigma}^2_{\hat{\beta}_1}$ 和 $\hat{\sigma}_{\hat{\beta}_0}{}^2$ 会产生有效的统计推断。因此不论误差项是否是异方差的，基于这些标准误的假设检验和置信区间都是有效的。因为不论误差项是否是异方差的，到目前为止，我们所使用的标准误（即基于公式(4.19)和公式(4.59)的那些标准误）都会导致有效的统计推断，所以它们被称为异方差稳健的标准误（heteroskedasticity－robust standard errors）。因为这些公式是由 Eicker(1967)、Huber(1967)和 White(1980)提出来的，所以它们也被称为 Eicker－Huber－White 标准误。

4.9.3 异方差和同方差在实际中意味着什么

就异方差和同方差来说，哪一个更符合实际？回答这个问题取决于实际应用情况。不过，我们可以回到不同性别的大学毕业生收入差距这个例子，用这个例子来阐明这个问题。至于哪一个假设更敏感，我们可以看一看周围现实世界里人们的工资怎样被支付，这可能会给我们提供一些思考的线索。许多年以来，我们很少会发现在一流报酬水平的工作中有女性：总是有报酬很低的男性，但几乎没有报酬很高的女性。这说明女性收入的分布要比男性收入的分布更紧凑。换句话说，公式(4.42)中女性的误差项的方差，似乎合理地低于公式(4.43)中男性的误差项的方差。因此，女性的工作和工资水平的这种"封顶现象"的存在，表明了公式(4.41)中二元变量回归模型的误差项是异方差的。除非存在无法抗拒的相反的原因（我们认为不会存在这些原因），否则，在这个例子中，把误差项作为异方差来处理是很有意义的。

上述这个对收入建模的例子表明，异方差在许多经济计量应用中都会出现。在一般情况下，经济理论很少会给出任何原因让人相信误差项是同方差的。因此，对误差项做出异方差的假设，这一做法是审慎的，除非你有无法拒绝的理由确信误差项是同方差的。

实际意义。这个讨论给我们引出的核心问题是，在实际中我们是应该使用异方差稳健的标准误还是使用仅适用于同方差的标准误？在这个问题上，想象一下，我们两个都计算，然后在它们之间进行选择，这是一种有用的做法。如果仅适用于同方差的标准误和异方差稳健的标准误是相同的，那么使用异方差稳健的标准误不会有任何损失；但如果它们不同，那么你就应该使用更可靠的即考虑到异方差的那一个。最简单的做法就是总是使用异方差



稳健的标准误。

由于历史原因，许多软件程序都用仅适用于同方差的标准误作为默认设置，因此是否使用异方差稳健的标准误这个选项，需要由使用者自己决定。如何去执行异方差稳健的标准误的计算细节，取决于你所使用的软件包。

在本书的所有经验例子中，除非有明确说明，否则我们都使用异方差稳健的标准误①。

4.10 结论

现在回到本章开头所提到的问题中，教育主管正考虑雇佣更多的教师来削减学生—教师比。到现在为止，我们学到了什么知识而且对她可能是有用的？

根据1998年加利福尼亚州的考试成绩数据集中的420个观测值，我们的回归分析表明，在学生—教师比和考试成绩之间存在负向的关系，即班级规模较小的地区会有较高的考试成绩。从实际意义上说，回归系数是比较大的：平均来看，学生—教师比每减少2个单位，其考试成绩平均就会高出4.6分。这相应地把该地区从考试成绩分布的第50个百分位数的位置，上移到大约第60个百分位数的位置。

学生—教师比这一回归系数统计上在5%的显著性水平下显著地异于0。总体系数可能是0，我们可能只用了一个随机的样本估计出了这个负的系数。然而，在其余的潜在的随机样本中通过纯粹随机抽样的方法估计出总体系数是0（以及获得一个和我们所得到的一样大的β_1的t统计量）的概率极其小，约为0.001%。β_1的95%的置信区间是$-3.30 \leqslant \beta_1 \leqslant -1.26$。

对于回答教育主管的问题，我们已取得了相当大的进展。然而，恼人的问题仍然存在。虽然我们估计了学生—教师比和考试成绩之间的负向关系，但是这种关系是否一定是教育主管进行决策所需要的那个因果关系呢？我们已发现平均来看，学生—教师比较低的地区具有较高的考试成绩，但这是否意味着降低学生—教师比确实能够提高成绩呢？

事实上，人们有理由担心降低学生—教师比可能不是提高学生成绩的原因。毕竟雇佣更多的教师需要花费资金，所以比较富裕的学区能够更好地负担得起较小的班级规模。但比较富裕的学校的学生还在其他方面而优于较穷的邻校学生，包括较好的设备、较新的课本和薪水较高的教师。另外，这些学校里的学生本身倾向于来自更富裕的家庭，从而拥有与他们的学校没有直接关系的其他优势。例如，加利福尼亚州有一个大规模的移民社区，这些移民一般比总体上的居民穷。在许多情况下，他们孩子的母语也不是英语。这样，在学生—教师比和考试成绩之间我们所估计出的负向关系，可能是由我们所发现的小班型连同许多其他因素一起作用的结果，而实际上这些其他因素可能是降低考试成绩的真正原因。

这些其他因素或"遗漏变量"可能意味着，迄今为止我们所做的OLS分析实际上对教育主管几乎没有价值。事实上它可能会造成误导：只改变学生—教师比，而不改变那些真正决定孩子在校成绩的其他因素。为了解决这个问题，我们需要一种在保持那些其他因素不变的情况下，允许我们将改变学生—教师比对考试成绩的影响分离出来的方法。这种方法就是多元回归分析，这是第5章的主题。



① 如果将本书和其他课本一起使用，请注意，一些课本把同方差加入到最小二乘假设中。然而，正如此处所讨论的，对OLS分析的有效性而言，这个额外的假设不是必须的，只要我们使用了异方差稳健的标准误。

总结

1. 总体回归线 $\beta_0+\beta_1X$ 是 Y 的均值(它是 X 值的函数)。斜率 β_1 是与 X 的一个单位变化相联系的 Y 的期望变化。截距项 β_0 决定回归线的水平(或高度)。重要概念 4.1 中总结了总体线性回归模型的术语。

2. 总体回归线可以采用普通最小二乘法(OLS),利用样本观测值 $(X_i,Y_i)(i=1,\cdots,n)$ 进行估计。回归直线的截距和斜率的 OLS 估计量分别表示为 $\hat{\beta}_0$ 和 $\hat{\beta}_1$。

3. 线性回归模型有三个重要的假设:(1)回归误差项 u_i 以回归因子 X_i 为条件的条件均值为零;(2)样本观测值是独立同分布的从总体中随机抽取的样本;(3)随机变量具有四阶矩。如果这些假设成立,那么 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是:(1)无偏的;(2)一致的;(3)当样本很大时,服从正态分布。

4. 回归系数的假设检验类似于总体均值的假设检验:使用 t 统计量计算 p 值,然后决定是接受还是拒绝零假设。和总体均值的置信区间估计方法一样,回归系数的 95% 的置信区间是估计量 ±1.96 倍标准误。

5. 当 X 是二元变量时,回归模型可被用来估计和检验关于"$X=0$"组和"$X=1$"组总体均值之差的假设。

6. R^2 和回归标准误(SER)是测度 Y_i 的值与所估计的回归直线之间紧密程度的指标。R^2 的取值在 0 和 1 之间,较大的值表明 Y_i 的值更接近于回归直线。回归标准误是回归误差项的标准差的一个估计量。

7. 通常,误差项 u_i 是异方差的,也就是说,给定 X_i 的值,u_i 的方差 $\text{var}(u_i|X_i=x)$ 依赖于 x。一个特殊的情况是,误差项是同方差的,即 $\text{var}(u_i|X_i=x)$ 为常数。当误差项是异方差的时,仅适用于同方差的标准误不会产生有效的统计推断,但是异方差稳健的标准误却会产生有效的统计推断。

重要术语

一元线性回归模型　因变量　自变量　回归因子　总体截距和斜率　总体回归线　总体系数　总体回归函数　参数　误差项　普通最小二乘(OLS)估计量　OLS 回归线　预测值　残差　最小二乘假设　$\hat{\beta}_1$ 的标准误　t 统计量　p 值　β_1 的置信区间　置信水平　指示变量　虚拟变量　乘以变量 D_i 的系数　D_i 的系数　回归的 R^2　被解释的平方和(ESS)　总平方和(TSS)　残差平方和(SSR)　回归标准误(SER)　异方差和同方差　最佳线性无偏估计量(BLUE)　加权最小二乘　仅适用于同方差的标准误　异方差稳健的标准误

复习概念

4.1　解释说明 $\hat{\beta}_1$ 与 β_1 之间的区别;残差 $\hat{u}_i$ 与回归误差 u_i 之间的区别;OLS 预测值 $\hat{Y}_i$ 和 $E(Y_i|X_i)$ 之间的区别。

4.2　简述利用独立同分布的 $Y_i(i=1,\cdots,n)$ 的观测值集合,计算双边检验 $H_0:\mu_Y=0$ 的 p 值的过程。简述在回归模型中利用独立同分布的观测值集合 $(X_i,Y_i)(i=1,\cdots,n)$,计算



双边检验 $H_0:\beta_1=0$ 的 p 值的过程。

4.3 解释说明如何利用3.5节中的数据,用回归模型来估计工资的性别差异。因变量和自变量分别是什么?

4.4 分别画出所估计的回归方程其拟合优度分别为 $R^2=0.9$ 和 $R^2=0.5$ 的假想的回归数据的散点图。



练习

标有*的习题解答可在本书的网址 www. aw. com/stock _ watson 上找到。

*4.1 假设研究人员利用来自100个三年级的班级的数据,研究了关于班级规模(CS)和平均考试成绩之间的关系,估计了OLS回归方程:

$$\widehat{TestScore} = \underset{(20.4)}{520.4} - \underset{(2.21)}{5.82} \times CS, R^2 = 0.08, SER = 11.5$$

a. 一个班有22个学生,这个班平均考试成绩的回归预侧值是多少?

b. 某班去年有19个学生,而今年有23个学生,这个班平均考试成绩变化的回归预测值是多少?

c. 构造回归斜率系数 β_1 的95%的置信区间。

d. 计算零假设 $H_0:\beta_1=0$ 的双边检验的 p 值。在5%的显著性水平下是否拒绝零假设?在1%的显著性水平下呢?

e. 在100个班级之间的样本平均班级规模为21.4,这100个班级之间的考试成绩的样本均值是多少?(提示:复习OLS估计量公式)

f. 这100个班级考试成绩的样本标准差是多少?(提示:复习 R^2 和 SER 的计算公式)

4.2 假设一位研究人员使用随机选择的250个男工人和280个女工人的工资数据,估计出OLS回归方程如下:

$$\widehat{Wage} = \underset{(0.18)}{12.68} + \underset{(0.84)}{2.79}\ Male, R^2 = 0.06, SER = 3.10$$

其中,$Wage$ 用美元/小时测度;$Male$ 是个二元变量,如果是男工人,那么 $Male=1$,如果是女工人,那么 $Male=0$。定义工资的性别差距为男女平均收人之差。

a. 所估计的(工资的)性别差距是多少?

b. 所估计的(工资的)性别差距是否显著地异于0(计算检验零假设"不存在性别差距"时所用的 p 值)?

c. 构造这个(工资的)性别差距的95%的置信区间。

d. 在这个样本中,女性的平均工资是多少?男性的平均工资是多少?

e. 另一位研究人员使用同样的数据,但是用 $Wage$ 对 $Female$ 进行回归,如果是女工人,则 $Female=1$;如果是男工人,则 $Female=0$。由这个回归所计算的回归估计值是多少?



$$\widehat{Wage} = _____ + _____ Female, R^2 = _____, SER = _____$$

*4.3 证明:第一个最小二乘假设 $E(u_i|X_i)=0$ 隐含 $E(Y_i|X_i)=\beta_0+\beta_1X_i$。



4.4 证明:$\hat{\beta}_0$ 是 β_0 的无偏估计量。(提示:利用 $\hat{\beta}_1$ 是无偏的事实,证明见附录4.3)

4.5 假设从总体中选择了200个20岁男性的随机样本,并记录了他们的身高和体重。体重对身高的回归方程为:

$$\widehat{Weight} = \underset{(2.15)}{-99.41} + \underset{(0.31)}{3.94}\ Height, R^2 = 0.81, SER = 10.2$$

其中,*Weight* 用磅测度,*Height* 用英寸测度。

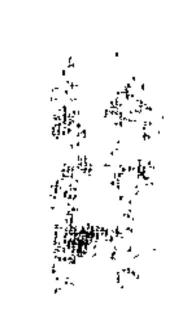
a. 对一个身高为70英寸的人而言,回归的体重预测值是多少?身高为60英寸和74英寸的人的体重呢?

b. 一个人发育较晚,并在一年内长高了1.5英寸。这个人的体重增加量的回归预测值是多少?

c. 构造(*b*)中体重增量的99%的置信区间。

d. 假设分别用千克和厘米来测度身高和体重,而不用磅和英寸,根据这个新的千克—厘米数据回归,回归估计值是多少(给出所有的结果、估计系数、标准误、R^2 和 SER)?

4.6 从公式(4.15)出发,推导出在附录4.4中公式(4.61)所给出的同方差条件下 $\hat{\beta}_0$ 的方差。





附录4.1 加利福尼亚州考试成绩数据集

加利福尼亚州标准化考试及报告的数据集包括考试成绩、学校特征和学生的人口统计背景的数据。这里所用的数据取自于所有在1998年和1999年可获得数据的加利福尼亚州的420个K-6和K-8地区。考试成绩是斯坦福9类达标考试(the Stanford 9 Achievement Test)(一个对五年级学生进行的标准化考试)中关于阅读和数学成绩的平均数。学校特征(地区间的平均)包括注册人数、教师的数量(用"全职教师等同量折合")、每间教室的计算机数量和每个学生的费用支出。这里所用的学生—教师比是该地区的全职教师等同量人数除以学生人数。学生的人口统计变量也是在地区间进行平均的。学生的人口统计变量包括:在公共援助计划CalWork(以前叫AFDC)中受援助学生的百分比、享有一份减价午餐的学生的百分比,以及学习英语的学生(即英语是其第二语言的学生)的百分比。所有这些数据都是从加利福尼亚州教育局得到的(www.cde.ca.gov)。

附录4.2 OLS估计量的推导

本附录利用微积分学知识推导了重要概念4.2中所给出的OLS估计量的公式。为了使预测误差的平方和 $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ (公式(4.6))最小化,首先分别对 b_0 和 b_1 求偏导数:

$$\frac{\partial}{\partial b_0}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i) \tag{4.44}$$

$$\frac{\partial}{\partial b_1}\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2 = -2\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)X_i \tag{4.45}$$

OLS估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是使 $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ 最小的 b_0 和 b_1 的值,或者说,使公式(4.44)和公式(4.45)中的导数等于0的 b_0 和 b_1 的值。因此,令这两个导数等于0,合并同类项,再除以 n,这证明了OLS估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 一定满足以下这两个方程:

$$\bar{Y} - \hat{\beta}_0 - \hat{\beta}_1\bar{X} = 0 \tag{4.46}$$



$$\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\hat{\beta}_0\overline{X}-\hat{\beta}_1\frac{1}{n}\sum_{i=1}^{n}X_i^2=0 \tag{4.47}$$

解这两个方程求出 $\hat{\beta}_0$ 和 $\hat{\beta}_1$：

$$\hat{\beta}_1=\frac{\frac{1}{n}\sum_{i=1}^{n}X_iY_i-\overline{X}\,\overline{Y}}{\frac{1}{n}\sum_{i=1}^{n}X_i^2-(\overline{X})^2}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{\sum_{i=1}^{n}(X_i-\overline{X})^2} \tag{4.48}$$

$$\hat{\beta}_0=\overline{Y}-\hat{\beta}_1\overline{X} \tag{4.49}$$

公式(4.48)和公式(4.49)就是重要概念 4.2 中所给出的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的表达式。通过将公式(4.48)的分子和分母同除以 $n-1$ 就可得到公式 $\hat{\beta}_1=s_{XY}/s_X^2$。

附录 4.3 OLS 估计量的抽样分布

在本附录中，我们证明了在重要概念 4.4 中所给出的 OLS 估计量 $\hat{\beta}_1$ 是无偏的，并在大样本条件下，$\hat{\beta}_1$ 服从正态的抽样分布。

用回归因子和误差项表示的 $\hat{\beta}_1$

我们首先给出用回归因子和误差项表达的 $\hat{\beta}_1$ 的表达式。由于 $Y_i=\beta_0+\beta_1X_i+u_i$，$Y_i-\overline{Y}=\beta_1(X_i-\overline{X})+u_i-\overline{u}$，因此公式(4.48)中 $\hat{\beta}_1$ 公式的分子是：

$$\begin{aligned}\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})&=\sum_{i=1}^{n}(X_i-\overline{X})[\beta_1(X_i-\overline{X})+(u_i-\overline{u})]\\&=\beta_1\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}(X_i-\overline{X})(u_i-\overline{u})\end{aligned} \tag{4.50}$$

现有 $\sum_{i=1}^{n}(X_i-\overline{X})(u_i-\overline{u})=\sum_{i=1}^{n}(X_i-\overline{X})u_i-\sum_{i=1}^{n}(X_i-\overline{X})\overline{u}=\sum_{i=1}^{n}(X_i-\overline{X})u_i$，其中，最后一个等式根据 $\overline{X}$ 的定义得出，它意味着 $\sum_{i=1}^{n}(X_i-\overline{X})\overline{u}=(\sum_{i=1}^{n}X_i-n\overline{X})\overline{u}=0$。将 $\sum_{i=1}^{n}(X_i-\overline{X})(u_i-\overline{u})=\sum_{i=1}^{n}(X_i-\overline{X})u_i$ 代入到公式(4.50)的最后一个表达式中得出 $\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})=\beta_1\sum_{i=1}^{n}(X_i-\overline{X})^2+\sum_{i=1}^{n}(X_i-\overline{X})u_i$，接下来将这个表达式再代入到公式(4.48)中，得到：

$$\hat{\beta}_1=\beta_1+\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2} \tag{4.51}$$

证明 $\hat{\beta}_1$ 是无偏的

通过对公式(4.51)的两边取期望，可以得到 $\hat{\beta}_1$ 的期望值，即：

$$E(\hat{\beta}_1)=\beta_1+E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}\right]$$

$$= \beta_1 + E\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})E(u_i \mid X_1, \cdots, X_n)}{\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2}\right] = \beta_1 \quad (4.52)$$

其中，公式(4.52)中的第二个等式根据累期望法则(见2.3节)得出。根据第二个最小二乘假设，对除了第 i 个以外的所有观测值而言，u_i 独立分布于 X，所以 $E(u_i \mid X_1, \cdots, X_n) = E(u_i \mid X_i)$。但根据第一个最小二乘假设，$E(u_i \mid X_i) = 0$。因此，公式(4.52)中最后一项的分子为0，这样 $E(\hat{\beta}_1) = \beta_1$，也就是说，OLS估计量是无偏的。

OLS估计量的大样本正态分布

通过考虑公式(4.51)中的最后一项的方式，可以得到 $\hat{\beta}_1$ 的极限分布的大样本正态近似(见重要概念4.4)。

首先考虑这项的分子。由于 $\overline{X}$ 是一致的，如果样本容量很大，则 $\overline{X}$ 几乎等于 μ_X。因此，可以从很逼近的意义上来说，等式(4.51)的分子中的项是样本均值 $\bar{v}$，这里 $v_i = (X_i - \mu_X)u_i$。根据第一个最小二乘假设，v_i 具有零均值；根据第二个最小二乘假设，v_i 是独立同分布的。v_i 的方差是 $\sigma_v^2 = \text{var}[(X_i - \mu_X)u_i]$；根据第三个最小二乘假设，它是非零的且有限的。因此，$\bar{v}$ 满足中心极限定理(见重要概念2.7)的所有要求条件。从而，在大样本条件下，$\bar{v}/\sigma_{\bar{v}}$ 服从 $N(0,1)$ 分布，其中 $\sigma_{\bar{v}}^2 = \sigma_v^2/n$。因此，$\bar{v}$ 的分布很好地被 $N(0, \sigma_v^2/n)$ 分布逼近。

接下来考虑公式(4.51)中分母的表达式，这是 X 的样本方差(不过这里是被 n 除而不是被 $n-1$ 除，如果 n 很大，这是无关紧要的)。如3.2节中所讨论的(公式(3.8))，样本方差是总体方差的一致估计量，所以在大样本条件下，它任意地接近于 X 的总体方差。

把这两个结果结合起来，我们有以下结论：在大样本条件下，$\hat{\beta}_1 - \beta_1 \approx \bar{v}/\text{var}(X_i)$，因此，$\hat{\beta}_1$ 的抽样分布在大样本条件下是 $N(\beta_1, \sigma_{\hat{\beta}_1}^2)$，这里 $\sigma_{\hat{\beta}_1}^2 = \text{var}(\bar{v})/[\text{var}(X_i)]^2 = \text{var}[(X_i - \mu_X)u_i]/\{n[\text{var}(X_i)]^2\}$，这就是公式(4.14)中的表达式。

关于OLS估计量的其他的一些代数结果

OLS残差和预测值满足：

$$\frac{1}{n}\sum_{i=1}^{n}\hat{u}_i = 0 \quad (4.53)$$

$$\frac{1}{n}\sum_{i=1}^{n}\hat{Y}_i = \overline{Y} \quad (4.54)$$

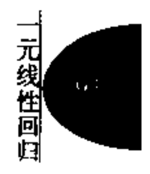
$$\sum_{i=1}^{n}\hat{u}_i X_i = 0, s_{\hat{u}X} = 0 \quad (4.55)$$

$$TSS = SSR + ESS \quad (4.56)$$

公式(4.53)到公式(4.56)说明了OLS残差的样本均值是0；OLS预测值的样本均值是 $\overline{Y}$；OLS残差和回归因子之间的样本协方差 $s_{\hat{u}X}$ 为0；总平方和是残差平方和与被解释的平方和之和(ESS，TSS 和 SSR 分别定义在公式(4.35)、公式(4.36)和公式(4.38)中)。

为了证明公式(4.53)，注意到 $\hat{\beta}_0$ 的定义使我们可将OLS残差写为 $\hat{u}_i = Y_i - \hat{\beta}_0 - \hat{\beta}_1 X_i = (Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})$，因此：

$$\sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}(Y_i - \overline{Y}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})$$



但 $\overline{Y}$ 和 $\overline{X}$ 的定义意味着 $\sum_{i=1}^{n}(Y_i - \overline{Y}) = 0$ 和 $\sum_{i=1}^{n}(X_i - \overline{X}) = 0$，所以，$\sum_{i=1}^{n}\hat{u}_i = 0$。

为了证明公式(4.54)，注意到有 $Y_i = \hat{Y}_i + \hat{u}_i$，所以 $\sum_{i=1}^{n} Y_i = \sum_{i=1}^{n}\hat{Y}_i + \sum_{i=1}^{n}\hat{u}_i = \sum_{i=1}^{n}\hat{Y}_i$，这里第二个等式是公式(4.53)的结果。

为了证明公式(4.55)，注意到 $\sum_{i=1}^{n}\hat{u}_i = 0$ 隐含着 $\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}\hat{u}_i(X_i - \overline{X})$，所以：

$$\sum_{i=1}^{n}\hat{u}_i X_i = \sum_{i=1}^{n}[(Y_i - \overline{Y}) - \hat{\beta}_1(X_i - \overline{X})](X_i - \overline{X})$$

$$= \sum_{i=1}^{n}(Y_i - \overline{Y})(X_i - \overline{X}) - \hat{\beta}_1\sum_{i=1}^{n}(X_i - \overline{X})^2 = 0 \tag{4.57}$$

其中，公式(4.57)中的最后一个等式是使用公式(4.48)中 $\hat{\beta}_1$ 的公式得到的。这个结果和前面的结果合并在一起隐含着 $s_{\hat{u}X} = 0$。

公式(4.56)可根据前面的结果和一些代数运算得到：

$$TSS = \sum_{i=1}^{n}(Y_i - \overline{Y})^2 = \sum_{i=1}^{n}(Y_i - \hat{Y}_i + \hat{Y}_i - \overline{Y})^2$$

$$= \sum_{i=1}^{n}(Y_i - \hat{Y}_i)^2 + \sum_{i=1}^{n}(\hat{Y}_i - \overline{Y})^2 + 2\sum_{i=1}^{n}(Y_i - \hat{Y}_i)(\hat{Y}_i - \overline{Y})$$

$$= SSR + ESS + 2\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = SSR + ESS \tag{4.58}$$

其中，最后一个等式是根据前面的结果 $\sum_{i=1}^{n}\hat{u}_i\hat{Y}_i = \sum_{i=1}^{n}\hat{u}_i(\hat{\beta}_0 + \hat{\beta}_1 X_i) = \hat{\beta}_0\sum_{i=1}^{n}\hat{u}_i + \hat{\beta}_1\sum_{i=1}^{n}\hat{u}_i X_i = 0$ 得到的。

附录 4.4　OLS 标准误的公式

本附录讨论 OLS 标准误的公式。这些公式最初是在重要概念 4.3 中的最小二乘假设下提出的，它们考虑到了异方差，是"异方差稳健的"标准误，然后给出了作为特殊情况的同方差的 OLS 估计量的方差，以及与之相联系的标准误公式。

异方差稳健的标准误

公式(4.19)中定义的估计量 $\hat{\sigma}^2_{\hat{\beta}_1}$，可用相应的样本方差代替公式(4.14)中的总体方差来获得，只是要做一些修正。公式(4.14)的分子中的方差是用 $\frac{1}{n-2}\sum_{i=1}^{n}(X_i - \overline{X})^2\hat{u}_i^2$ 估计的，其中除数 $n-2$（而不是 n）加入了一个自由度调整以修正向下的偏差，类似于 4.8 节 *SER* 定义中所使用的自由度调整。分母中的方差是用 $\frac{1}{n}\sum_{i=1}^{n}(X_i - \overline{X})^2$ 估计的。用这两个估计量代替公式(4.14)中的 $\mathrm{var}[(X_i - \mu_X)u_i]$ 和 $\mathrm{var}(X_i)$，便得到公式(4.19)中的 $\hat{\sigma}^2_{\hat{\beta}_1}$。异方差稳健的标准误的一致性问题在 15.3 节中讨论。

$\hat{\beta}_0$ 的方差估计量是：

$$\hat{\sigma}^2_{\hat{\beta}_0} = \frac{1}{n} \times \frac{\frac{1}{n-2}\sum_{i=1}^{n}\hat{H}_i^2\hat{u}_i^2}{\left(\frac{1}{n}\sum_{i=1}^{n}\hat{H}_i^2\right)^2} \tag{4.59}$$

其中，$\hat{H}_i = 1 - \left(\bar{X} \Big/ \frac{1}{n}\sum_{i=1}^{n} X_i^2\right) X_i$。$\hat{\beta}_0$ 的标准误是 $SE(\hat{\beta}_0) = \sqrt{\hat{\sigma}_{\hat{\beta}_0}^2}$。估计量 $\hat{\sigma}_{\hat{\beta}_0}^2$ 背后的推理和 $\hat{\sigma}_{\hat{\beta}_1}^2$ 的一样，源自于用样本均值代替总体期望值。

仅适用于同方差的方差

在同方差的情况下，给定 X_i 条件下 u_i 的条件方差是个常数，即 $\text{var}(u_i|X_i) = \sigma_u^2$。如果误差项是同方差的，那么重要概念4.4 中的公式可简化为：

$$\sigma_{\hat{\beta}_1}^2 = \frac{\sigma_u^2}{n\sigma_X^2} \tag{4.60}$$

$$\sigma_{\hat{\beta}_0}^2 = \frac{E(X_i^2)}{n\sigma_X^2}\sigma_u^2 \tag{4.61}$$

为了推导公式(4.60)，将公式(4.14)中的分子写为 $\text{var}[(X_i - \mu_X)u_i] = E(\{(X_i - \mu_X)u_i - E[(X_i - \mu_X)u_i]\}^2) = E\{[(X_i - \mu_X)u_i]^2\} = E[(X_i - \mu_X)^2 u_i^2] = E[(X_i - \mu_X)^2 \text{var}(u_i | X_i)]$，这里第二个等式是根据 $E[(X_i - \mu_X)u_i] = 0$ 得到的(根据第一个最小二乘假设)，而最后一个等式是根据累期望法则(见2.3 节)得到的。如果 u_i 是同方差的，那么 $\text{var}(u_i|X_i) = \sigma_u^2$，所以 $E[(X_i - \mu_X)^2 \text{var}(u_i|X_i)] = \sigma_u^2 E[(X_i - \mu_X)^2] = \sigma_u^2 \sigma_X^2$。将这个表达式代入公式(4.14)的分子中并化简，就得到公式(4.60)中的结果。类似的计算可得到公式(4.61)。

仅适用于同方差的标准误

用样本均值、方差分别代替公式(4.60)和公式(4.61)中的总体均值与方差，并用 SER 的平方来估计 u_i 的方差，就可得到仅适用于同方差的标准误。这些方差的仅适用于同方差的估计量为：

$$\tilde{\sigma}_{\hat{\beta}_1}^2 = \frac{s_{\hat{u}}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \text{（仅适用于同方差的）} \tag{4.62}$$

$$\tilde{\sigma}_{\hat{\beta}_0}^2 = \frac{\left(\frac{1}{n}\sum_{i=1}^{n} X_i^2\right) s_{\hat{u}}^2}{\sum_{i=1}^{n}(X_i - \bar{X})^2} \text{（仅适用于同方差的）} \tag{4.63}$$

其中，$s_{\hat{\beta}_0}^2$ 在公式(4.40)中给出。仅适用于同方差的标准误是 $\tilde{\sigma}_{\hat{\beta}_0}^2$ 和 $\tilde{\sigma}_{\hat{\beta}_1}^2$ 的平方根。

第5章 多元线性回归

第4章以一个使人困惑的评论结束。在加利福尼亚州数据集中，尽管学生—教师比较低的地区倾向于有较高的考试成绩，但来自于小班地区的学生也有可能拥有其他一些有助于他们在标准化考试中表现好的优势。这个问题是否令人费解？如果是，我们应该怎样解决呢？

我们在使用普通最小二乘法（OLS）估计出的班级规模对考试成绩影响效应的估计量中，由于遗漏了像学生特征这样的因素，因此会产生误导性的结论，或者更确切地说，产生偏差。本章中我们解释了这个“被遗漏的变量所导致的偏差”，并引入了多元回归方法，这种方法是一种能够削减遗漏变量偏差的方法。多元回归方法的关键思想是：如果我们有关于这些遗漏变量的数据，那么我们就能把它们作为额外的回归因子包括进来，从而在保持其他变量（如学生特征）不变的条件下估计单个回归因子（学生—教师比）的效应。

本章解释说明了如何估计多元线性回归模型的系数，仔细研究了如何进行统计推断，即如何检验关于多个回归系数的假设和如何构造那些系数的置信区间。多元回归的许多方面与第4章中所研究的一元回归相似。多元回归模型的系数也可以使用OLS方法根据样本数据进行估计；多元回归的OLS估计量也是随机变量，因为它们所依赖的数据也是从随机样本中得到的；在大样本中，OLS估计量的抽样分布也是渐近正态的；OLS估计量也可被用来对总体回归系数作假设检验和构造置信区间。这样一个可检验的假设是：在保持该地区可测度的学生特征不变的情况下，降低学生—教师比对考试成绩没有影响。

5.1 遗漏变量偏差

由于只研究学生—教师比这一变量，因此，第4章中的实证分析忽略了一些潜在的可能决定考试成绩的重要因素，而这些可能的重要因素的影响是通过回归误差项反映的。这些被遗漏的因素包括：学校特征，如教师素质和计算机的使用情况；学生特征，如家庭背景。我们首先考虑一个被遗漏的学生特征，即该校区中仍然在学习英语的学生的比例，这个指标之所以在加利福尼亚州特别重要是因为该州拥有大量的移民人口。

由于忽略了这个地区英语学习者的百分比，因此，我们所估计的考试成绩对学生—教师比的回归斜率的OLS估计量可能是有偏的，即OLS估计量的抽样分布的均值可能不等于学生—教师比的单位变化对考试成绩影响的真实效应。我们做以下推理：仍在学英语的学生在标准化考试中可能会比把英语作为母语的学生表现得差。如果大班型的地区中学习英语的学生也很多，那么考试成绩对学生—教师比的OLS回归可能会得到错误的相关关系，并得到一个较大的估计系数，而削减班级规模对考试成绩的真实因果效应实际上可能是很小的，甚至是零。因此，根据第4章的分析，该教育主管可能会雇佣足够多的新教师以使学生—教师比减少2个单位，但是如果真实的系数很小或是零，那么她原本希望考试成绩提高的愿望将不会实现。



看一下加利福尼亚州的数据就可为这一点提供证据。该地区学生—教师比和英语学习者（即那些英语不是母语的学生以及那些还未掌握英语的学生）的百分比之间的相关系数为0.19。这个很小但却正的相关系数表明，英语学习者越多的地区倾向于有更高的学生—教师比（较大的班级规模）。如果学生—教师比与英语学习者的百分比不相关，那么在考试成绩对学生—教师比的回归分析中，忽略英语熟练程度这一因素将是安全的。但由于学生—教师比与英语学习者的百分比之间是相关的，因此在考试成绩对学生—教师比的回归中，所得出的OLS系数就有可能反映了这种影响。

5.1.1 遗漏变量偏差的定义

如果回归因子（学生—教师比）与回归分析中已遗漏的某个变量（英语学习者的百分比）相关，而且该变量部分地决定因变量（考试成绩）的变化，那么该OLS估计量将会存在**遗漏变量偏差（omitted variable bias）**。

如果下面两个条件成立，那么就会发生遗漏变量偏差：第一，遗漏变量与包含在回归方程中的回归因子相关；第二，遗漏变量是因变量的一个决定因素。为了举例说明这些条件，我们用三个例子来说明考试成绩对学生—教师比回归中遗漏变量的问题。

例1：英语学习者的百分比。由于英语学习者的百分比与学生—教师比是相关的，因此遗漏变量偏差的第一个条件成立。那些仍在学习英语的学生在标准化考试中的表现要比说英语母语的学生差似乎是合理的，在这种情况下，英语学习者的百分比是影响考试成绩的一个决定性因素，从而遗漏变量偏差的第二个条件成立。因此，考试成绩对学生—教师比回归中的OLS估计量可能错误地反映了遗漏变量即英语学习者的百分比的影响。也就是说，遗漏英语学习者的百分比这一变量可能会引致遗漏变量偏差。

例2：安排考试的时间。分析中遗漏的另一个变量是安排考试的时间。对这个遗漏变量而言，遗漏变量偏差的第一个条件不成立，但第二个条件成立，这似乎是合理的。例如，如果考试时间的安排以一种与班级规模不相关的方式在地区间变化，那么考试时间安排与班级规模就会不相关，所以第一个条件也就不成立。另一方面，考试时间的安排可能会影响考试成绩（学生思维的敏捷性在一天中的不同时间是不同的），因此第二个条件成立。然而，在这个例子中，由于安排考试的时间和学生—教师比是不相关的，因此学生—教师比不会错误地把“日间效应”反映进来。所以这里遗漏考试时间安排这一因素不会导致遗漏变量偏差。

例3：平均每个学生的停车场空间。另一个遗漏变量是每个学生的停车场空间（教师停车场面积除以学生数）。这个变量满足遗漏变量偏差的第一个条件，但不满足第二个条件。具体地讲，每个学生对应较多的老师的学校可能会有较大的教师停车场空间，所以第一个条

件会被满足。然而，在"学习是在教室里而不是在停车场上进行的"这一假设下，停车场空间对学习没有直接影响，因而第二个条件不成立。因为每位学生的停车场空间不是考试成绩的决定性因素，所以分析中遗漏此变量不会导致遗漏变量偏差。

关于遗漏变量偏差的概要解释，总结在重要概念 5.1 中。

遗漏变量偏差和第一个最小二乘假设。遗漏变量偏差意味着第一个最小二乘假设——即 $E(u_i|X_i)=0$，如在重要概念 4.3 中所列出的——是不正确的。为了理解其原因，回想一下，一元线性回归模型中的误差项 u_i 代表了除 X_i 之外决定 Y_i 的所有其他因素。如果这些其他因素中有一个与 X_i 相关，那么这就意味着误差项 u_i（它包含这个因素）与 X_i 是相关的。换句话说，如果遗漏变量是 Y_i 的一个决定性因素，那么它就在误差项中，而如果它与 X_i 相关，那么误差项就与 X_i 相关。由于 u_i 与 X_i 是相关的，因此给定 X_i 条件下 u_i 的条件均值就是非零的。所以，这个相关性违背了第一个最小二乘假设，其后果是严重的：OLS 的估计量是有偏的。这个偏差即使在非常大的样本中也不会消失，而且 OLS 的估计量也是不一致的。

5.1.2 遗漏变量偏差的计算公式

前一部分关于遗漏变量偏差的讨论在数学上可用一个公式对其进行概括。令 X_i 与 u_i 之间的相关系数为 $\mathrm{corr}(X_i,u_i)=\rho_{Xu}$。假定最小二乘假设的第二个和第三个假设成立，但第一个假设不成立，因为 ρ_{Xu} 是非零的，那么 OLS 估计量有极限（推导见附录 5.1），即：

$$\hat{\beta}_1 \xrightarrow{p} \beta_1 + \rho_{Xu}\frac{\sigma_u}{\sigma_X} \tag{5.1}$$

也就是说，随着样本容量的增加，$\hat{\beta}_1$ 以逐渐增高的概率逼近于 $\beta_1+\rho_{Xu}\frac{\sigma_u}{\sigma_X}$。

重要概念 5.1

一元回归中的遗漏变量偏差

遗漏变量偏差是指当回归因子 X 与遗漏变量相关时 OLS 估计量中出现的偏差。要发生遗漏变量偏差，两个条件必须为真：

1. X 与遗漏变量相关；

2. 遗漏变量是因变量 Y 的一个决定性因素。

表达式(5.1)中的公式概括了上面讨论的关于遗漏变量偏差的几个思想：

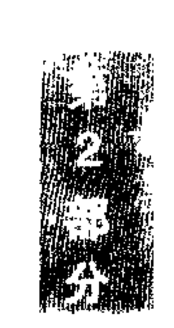
1. 遗漏变量偏差是一个不论样本容量是大还是小都会存在的问题。由于 $\hat{\beta}_1$ 并不依概率收敛于真实值 β_1，因此 $\hat{\beta}_1$ 不是一致性的。也就是说，当存在遗漏变量偏差时，$\hat{\beta}_1$ 不是 β_1 的一致性估计量。表达式(5.1)中的 $\rho_{Xu}\frac{\sigma_u}{\sigma_X}$ 项就是即使在大样本条件下依然存在的 $\hat{\beta}_1$ 的偏差。

2. 在实际中，这个偏差是大还是小，依赖于回归因子和误差项之间的相关系数 ρ_{Xu}。$|\rho_{Xu}|$ 越大，偏差就越大。

3. $\hat{\beta}_1$ 中偏差的方向依赖于 X 和 u 之间是正相关还是负相关。例如，我们推测学习英语的学生百分比对地区考试成绩有负效应（仍在学英语的学生有较低的分数），所以英语学习者的百分比以负号进入误差项。在我们的数据中，由于英语学习者的百分比与学生一教师

比是正相关的(有较多英语学习者的地区有较大的班级规模),因此,学生—教师比(X)与误差项(u)将会是负相关的,即$\rho_{Xu}<0$,学生—教师比的系数$\hat{\beta}_1$将会偏向于负数。换句话说,具有较小的英语学习者百分比,既与高的考试成绩相关联,也与低的学生—教师比相关联。所以,OLS 估计量显示出小的班级规模改进了考试成绩,导致这一结论的一个原因可能是小班地区有较少的英语学习者。

5.1.3 通过将数据分组来处理遗漏变量偏差



对于遗漏变量偏差,你能做些什么?我们的教学主管正考虑增加其所在地区的教师数量,但是她无法控制其所在社区中移民人数的百分比。其结果是,她对学生—教师比对考试成绩的影响感兴趣,而保持其他的因素不变,包括英语学习者的百分比。现在我们用一种新的方式提出了教育主管的问题,这种方式表明,我们可能应该集中关注那些英语学习者的百分比可以和该教育主管所在地区该指标具有可比性的地区,而不是所有地区的数据。在这些地区的数据子集中,那些规模较小班级的地区在标准化考试中会做得较好吗?

表 5—1 报告了英语学习者的百分比具有可比性的地区班级规模和考试成绩之间关系的证据。所有地区被分成 8 组。首先,对应于各个地区的英语学习者百分比数据分布的四分位数,将这些地区分成四类;然后,在这四类地区中,每一类地区又进一步被划分为两组,这取决于学生—教师比是小的($STR<20$)还是大的($STR\geq20$)。

表 5—1　**加利福尼亚州学区考试成绩的差异与高低学生—教师比数据**

(按该地区英语学习者的百分比进行分组)

	学生—教师比 <20		学生—教师比 ≥20		高低 STR 的考试成绩差异	
	平均考试分数	n	平均考试分数	n	差异	t 统计量
所有地区	657.4	238	650.0	182	7.4	4.04
英语学习者的百分比						
<2.2%	664.1	78	665.4	27	-1.3	-0.44
2.2% ~8.8%	666.1	61	661.8	44	4.3	1.44
8.8% ~23.0%	654.6	55	649.7	50	4.9	1.64
>23.0%	636.7	44	634.8	61	1.9	0.68

一般兴趣框

莫扎特(Mozart)效应:存在遗漏变量偏差吗

1993 年《自然》杂志上发表的一篇论文(Rauscher, Shaw, Ky, 1993)提出,听莫扎特的曲子 10~15 分钟会暂时使你的 IQ 提高 8 或 9 分。该项研究引起了很大的轰动——政治家们和父母们看到了使他们的孩子变得更聪明的一种容易的方法。有一段时间里,佐治亚州甚至向该州所有婴儿分发古典音乐 CD。



"莫扎特效应"的证据是什么?某人在对许多研究成果做出总结性研究后发现,那些在高中选修音乐或艺术课程的学生的英语和数学考试成绩确实比那些没有选修该类课程的学生成绩高。[①]然而,更深入地分析这些研究成果你会发现,拥有较好的考试成绩的真正原因和那些课程几乎没有关系。相反,这篇总结性研究的作者提出,考得好和上艺术或音乐课之间的相关关系可能是由许多事情引起的。例如,学习较好的学生可能会有更多的时间去选修音乐课,或对选修这些课程有更多的兴趣,或那些有较深的音乐课程的学校恰好是各学校

中较好的学校。

用回归术语来说，所估计的考试成绩和选修音乐课程之间的关系看起来似乎存在遗漏变量偏差。由于遗漏了因素，诸如学生天赋或学校整体质量，学习音乐表面上看对考试成绩有影响，但实际上它对考试成绩并没有影响。

那么，莫扎特效应到底存在不存在？找出结论的一种方法是进行随机化控制实验（如在第11章进一步讨论的，随机化控制实验通过将参与者随机地分配给"处理"组和"控制"组来剔除遗漏变量偏差）。把所有研究结果都放到一起，发现许多莫扎特效应的控制实验没有表明听莫扎特的曲子会提高IQ或提高通常的考试成绩。然而，由于某些无法完全理解的原因，听古典音乐在一个狭窄的领域内确实暂时地有帮助：折纸和想像形状。所以下次你在为折纸手工考试而死记硬背时，也试着准备听一些莫扎特的曲子。

注：①*Journal of Aesthetic Education* 34：3－4（Fall/Winter 2000），especially the article by Ellen Winner and Monica Cooper（pp. 11－76）and the one by Lois Hetland（pp. 105－148）。

表5—1中的第一行给出了高低学生—教师比地区之间平均考试成绩的总体差异，即在没有将其进一步分成英语学习者的四分位数的条件下，这两组考试成绩之间的差异。（回想一下，这个差异在前面的回归方程(4.33)中就被报告过，它是在对 D_i 的 *TestScore* 的回归方程中作为 D_i 系数的OLS估计值被报告出来，其中 D_i 是个二元回归因子，若 $STR_i<20$，则 D_i 等于1，否则 D_i 等于0）在420个地区的全样本中，低学生—教师比地区的平均考试成绩比高学生—教师比地区的考试成绩高7.4分，t 统计量为4.04，所以这两组平均考试成绩是相同的零假设在1%的显著水平下被拒绝。

表5—1中的后四行给出了按照英语学习者百分比数据分布的四分位数来划分的高低学生—教师比地区之间的平均考试成绩差异值。这个证据显示了一种不同的结论。在英语学习者最少（<2.2%）的地区中，78个低学生—教师比地区的平均考试成绩为664.1，而27个高学生—教师比地区的平均考试成绩为665.4。因此，对英语学习者较少的地区而言，具有较低学生—教师比地区的考试成绩比具有较高学生—教师比地区的考试成绩平均看来要低1.3分；在第二个四分位数中，低学生—教师比地区比高学生—教师比地区的考试成绩平均高出4.3分；对第三个四分位数而言，这个差距是4.9分；而对英语学习者最多的四分位数的地区而言，这个差距只有1.9分。一旦我们保持英语学习者百分比这个指标不变，高低学生—教师比地区之间的考试成绩的差异可能是这个总体估计值7.4分的一半（或更少）。

起初这个发现可能令人困惑。考试成绩的整体效应怎么可能是任何一个四分位数考试成绩效应的两倍呢？答案是英语学习者最多的地区倾向于有最高的学生—教师比和最低的考试成绩。最低和最高英语学习者百分比的四分位数地区之间的平均考试成绩差异很大，约为30分。英语学习者少的地区倾向于有较低的学生—教师比：在英语学习者的第一个四分位数内74%（78/105）的地区有小班（$STR<20$），而在英语学习者最多的四分位数内只有42%（44/105）的地区有小班。所以，英语学习者最多的地区有比其他地区低的考试成绩，还有比其他地区高的学生—教师比。

这个分析加大了教育主管的烦恼，即在考试成绩对学生—教师比的回归分析中存在遗漏变量偏差。通过在英语学习者百分比的四分位数内进行观察，表5—1第二部分中的考试成绩差异比表5—1第一行中表现的简单均值差有所改进。尽管如此，在保持英语学习者的百分比不变的条件下，这个分析仍然没有为教育主管提供一个改变班级规模对考试成绩的影响的有用估计值。不过，使用多元回归的方法能够提供一个这样的估计值。

而，正如一元回归中的情况一样，这个关系不会精确地成立，因为有许多其他因素影响着因变量。例如，除了学生—教师比和仍在学习英语的学生百分比外，考试成绩还受到学校特征、其他学生特征和学生运气的影响，因此，需要将公式(5.2)中的总体回归函数扩大，以加入这些额外的因素。

正如一元回归中的情况一样，除了 X_{1i} 和 X_{2i} 之外，决定 Y_i 的因素都被当做误差项 u_i 加入到公式(5.2)中了。这个误差项是特定的观察值(在我们的例子中为第 i 个地区的考试成绩)与平均总体关系的偏差，因此，我们有：



$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, i = 1, \cdots, n \tag{5.5}$$

这里，下标 i 代表样本中 n 个观察值(地区)中的第 i 个。

公式(5.5)是当有两个回归因子 X_{1i} 和 X_{2i} 时的**总体多元回归模型**(population multiple regression model)。

在含有二元回归因子的回归中，将 β_0 看做为恒等于 1 的回归因子的系数是很有用的，可将 β_0 看做是 X_{0i} 的系数，这里，对于 $i=1,\cdots,n$，有 $X_{0i}=1$。因此，公式(5.5)中的总体多元回归模型可改写成：

$$Y_i = \beta_0 X_{0i} + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i, \text{这里 } X_{0i} = 1, i = 1, \cdots, n \tag{5.6}$$

公式(5.5)和公式(5.6)这两种表示总体多元回归模型的方法是等价的。

到目前为止，我们的讨论一直集中于额外增加一个变量 X_2 的情形。然而实际上，可能有多个因素被单个回归因子模型遗漏掉了。例如，就像忽略英语学习者的百分比一样，忽略学生的经济背景也可能会导致遗漏变量偏差。这个推理引导我们考虑一个含有三个回归因子的模型，或更一般地说，含有 k 个回归因子的模型。含有 k 个回归因子 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 的多元回归模型在重要概念 5.2 中总结。

多元回归模型中的同方差和异方差的定义类似于单个回归因子模型中同方差和异方差的定义。在多元回归模型中，如果给定 $X_{1i}, X_{2i}, \cdots, X_{ki}$，$u_i$ 的条件分布的方差 $\text{var}(u_i | X_{1i}, X_{2i}, \cdots, X_{ki})$ 对于所有的 $i=1,\cdots,n$ 都是不变的，进而不依赖于 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 的值，那么误差项 u_i 就是**同方差的**(homoskedastic)，否则，误差项就是**异方差的**(heteroskedastic)。

多元回归模型可以解决前面的教育主管所要解决的问题，即改变学生—教师比对考试成绩的影响，同时对她所不能控制的因素保持不变，这些因素不仅包含英语学习者的百分比，还包含其他可能会影响考试成绩的可测度的因素，如学生的经济背景。然而，为了给教育主管提供实际的帮助，我们需要为她提供利用样本数据计算的总体回归模型的未知总体系数 $\beta_0, \beta_1, \cdots, \beta_k$ 的估计值。幸运的是，这些系数可用 OLS 进行估计。

重要概念 5.2

多元回归模型

多元回归模型是：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \cdots, n \tag{5.7}$$

其中：

- Y_i 是因变量的第 i 个观察值；$X_{1i}, X_{2i}, \cdots, X_{ki}$ 是 k 个回归因子中每一个的第 i 个观察值；u_i 是误差项。
- 总体回归线是总体中平均来看在 Y 和多个 X 之间成立的关系：

$$E(Y | X_{1i} = x_1, X_{2i} = x_2, \cdots, X_{ki} = x_k) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$$

- β_1 是 X_1 的斜率系数，β_2 是 X_2 的斜率系数，依此类推。系数 β_1 是在 $X_{2i}, \cdots, X_{ki}$ 保持

不变的情况下，由 X_{1i} 单位变化所引起的 Y_i 的期望变化。其他 X 变量的系数也可类似地进行解释。

• 截距 β_0 是当所有的 X 都为零时 Y 的期望值，截距可被看做是回归因子 X_{0i} 的系数，对于所有的 i 它都等于 1。

5.3 多元回归中的 OLS 估计量



本节描述了如何使用 OLS 估计多元回归模型的系数。

5.3.1 OLS 估计量

在 4.2 节中，我们说明了如何根据 Y 和 X 的一个样本观察值，运用 OLS 方法估计单个回归因子模型中的截距和斜率系数，其中关键的思想是，通过最小化预测误差平方和，即通过选择估计量 b_0 和 b_1 使得 $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_i)^2$ 的值最小来估计这些系数，这样得到的估计量就是 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$。

OLS 方法也可用来估计多元回归模型中的系数 $\beta_0,\beta_1,\cdots,\beta_k$。令 $b_0,b_1,\cdots,b_k$ 为 $\beta_0,\beta_1,\cdots,\beta_k$ 的估计量，使用这些估计量所计算的 Y_i 的预测值是 $b_0 + b_1X_{1i} + \cdots + b_kX_{ki}$，$Y_i$ 的预测误差是 $Y_i - (b_0 + b_1X_{1i} + \cdots + b_kX_{ki}) = Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki}$。因而，对于所有的 n 个观察值，这些预测误差的平方和为：

$$\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2 \tag{5.8}$$

表达式(5.8)中的线性回归模型的误差平方和是表达式(4.6)中所给出的单个回归因子线性回归模型的误差平方和的扩展形式。

使表达式(5.8)的误差平方和最小的系数 $\beta_0,\beta_1,\cdots,\beta_k$ 的估计量被称为 $\beta_0,\beta_1,\cdots,\beta_k$ 的**普通最小二乘(OLS)估计量(ordinary least squares estimator)**。OLS 估计量表示为 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$。

多元线性回归模型中 OLS 的术语与一元线性回归模型中的术语相同。OLS 回归线(OLS regression line)是使用 OLS 估计量构造的直线，即 $\hat{\beta}_0 + \hat{\beta}_1X_1 + \cdots + \hat{\beta}_kX_k$。根据 OLS 回归线，给定 $X_{1i},\cdots,X_{ki}$ 条件下 Y_i 的预测值(predicted value)是 $\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}$。第 i 个观察值的 OLS 残差(OLS residual)是 Y_i 与它的 OLS 预测值之差，即 OLS 残差 $u_i = Y_i - \hat{Y}_i$。

OLS 估计量可以通过反复尝试不同的 $b_0,\cdots,b_k$ 的值，经试验计算，直到你满意地认为表达式(5.8)中的总平方和已达到最小为止。不过，使用微积分学知识导出的 OLS 估计量的显性表达式是更容易的。多元回归模型中的 OLS 估计量的表达式与重要概念 4.2 中单个回归因子模型中估计量的表达式类似。这些表达式被加入到现代统计软件中。在多元回归模型中，表达式最好用矩阵符号来表示和讨论，所以它们的介绍被推迟到 16.1 节。

多元回归中 OLS 的定义和术语在重要概念 5.3 中总结。

5.3.2 在考试成绩和学生—教师比案例中的应用

在 4.2 节，我们利用加利福尼亚州校区中的 420 个观察值，使用 OLS 法估计了考试成绩

(*TestScore*)与学生—教师比(*STR*)回归中的截距和斜率系数。在方程(4.7)中报告的所估计的 OLS 回归线是:

$$\widehat{TestScore} = -698.9 - 2.28 \times STR \tag{5.9}$$

我们一直在担心这个关系可能会产生误导,因为学生—教师比可能会忽略在大班中有很多英语学习者这一因素对考试成绩的影响,也就是说,OLS 估计量存在遗漏变量偏差。

现在我们正通过使用 OLS 方法估计多元回归模型来处理这个问题,在这个多元回归中,因变量是考试成绩(Y_i),并有两个回归因子:420 个地区($i=1,\cdots,420$)的学生—教师比(X_{1i})和校区中英语学习者的百分比(X_{2i})。根据这个多元回归模型所估计的 OLS 回归线为:



$$\widehat{TestScore} = 686.0 - 1.10 \times STR - 0.65 \times PctEL \tag{5.10}$$

这里,*PctEL* 是地区中学习英语学生的百分比。截距($\hat{\beta}_0$)的 OLS 估计值为 686.0,学生—教师比系数($\hat{\beta}_1$)的 OLS 估计值为 -1.10,英语学习者百分比系数($\hat{\beta}_2$)的 OLS 估计值为 -0.65。

在多元回归中,学生—教师比的变化对考试成绩的估计影响大约相当于学生—教师比是惟一回归因子时的一半:在单个回归因子的方程中(公式(5.9)),估计 *STR* 每减少 1 单位会使考试成绩提高 2.28 分,但在多元回归方程(公式(5.10))中,估计结果是只会使考试成绩提高 1.10 分。发生这种差异的原因是,在多元回归中,*STR* 的系数是在保持(或控制)*PcEL* 不变时,*STR* 的变化所带来的影响,而在单个回归因子的回归中,*PctEl* 不是保持不变的。

由于我们认为公式(5.9)中的单个回归因子模型的估计值可能存在遗漏变量偏差,因此,相比之下,现在得出的多元回归的这两个估计值更为协调一致。在 5.1 节,我们看到英语学习者百分比高的地区不仅倾向于有低的考试成绩,还有高的学生—教师比。如果回归中遗漏了英语学习者的百分比,则我们估计出的减少学生—教师比对考试成绩的影响会更大。但是,多元回归估计值既反映了学生—教师比的变化对考试成绩的影响,也反映了曾被我们遗漏的地区中有较少英语学习者对考试成绩的影响。

重要概念 5.3

多元回归模型中的 OLS 估计量、预测值和残差

OLS 估计量 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$ 是使预测误差平方和 $\sum_{i=1}^{n}(Y_i - b_0 - b_1X_{1i} - \cdots - b_kX_{ki})^2$ 最小的 $b_0,b_1,\cdots,b_k$ 的值。OLS 预测值 $\hat{Y}_i$ 和残差 $\hat{u}_i$ 分别为:

$$\hat{Y}_i = \hat{\beta}_0 + \hat{\beta}_1X_{1i} + \cdots + \hat{\beta}_kX_{ki}, i=1,\cdots,n \tag{5.11}$$

$$\hat{u}_i = Y_i - \hat{Y}_i, i=1,\cdots,n \tag{5.12}$$

OLS 估计量 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$ 和残差 $\hat{u}_i$ 是根据 n 个观察值的样本($X_{1i},\cdots,X_{ki},Y_i$),$i=1,\cdots,n$ 计算出来的,它们也是未知的真实总体系数 $\beta_0,\beta_1,\cdots,\beta_k$ 和误差项 u_i 的估计值。

通过两种不同的途径我们得出了同样的结论,即在考试成绩和学生—教师比之间的关系中存在遗漏变量偏差:将数据分组的列表方法(见 5.1 节)和多元回归方法(公式(5.10))。在这两种方法中,多元回归方法有两个重要的优点:第一,它提供了学生—教师比减少 1 个单位对考试成绩影响的定量估计值,这是教育主管决策时所需要的;第二,它易

间变化。如果除了 STR_i 和 $PctEL_i$ 之外，变量 $FracEL_i$ 被作为第三个回归因子包含进来，那么回归因子将会是完全多重共线的，因为 $PctEL$ 是英语学习者百分比。因此，对于每个地区，有 $PctEL_i = 100 \times FracEL_i$。这样，一个回归因子($PctEL$)可被表示成另一个回归因子($FracEL_i$)的精确的线性函数。

由于存在这种完全多重共线性，因此，计算 $TestScore_i$ 对 STR_i，$PctEL_i$ 和 $FracEL_i$ 的回归的 OLS 估计值是不可能的。不同的软件包对完全多重共线性的处理方式不同。如果你试图估计这个回归，那么软件将会输出下面三个结果中的一个：它会剔除一个变量（任意地选择一个剔除）；它会拒绝计算 OLS 估计值并给出错误信息；或者死机。之所以会这样的数学原因是，多重共线性在 OLS 公式中产生了 0 除数的情况。

直观上，完全多重共线性之所以是个难题的原因在于你总是用回归去解答一个不合逻辑的问题。回想一下，$PctEL_i$ 的系数是在保持其他变量不变的条件下，$PctEl$ 的单位变化对考试成绩的影响。如果其他变量之一是 $FracEl$，那么在保持英语学习者比例不变的条件下，英语学习者百分比的单位变化的影响是什么？由于在精确的线性关系中，英语学习者的百分比和英语学习者的比例一起变动，因此这个问题没有意义，OLS 也不能解答这个问题。

例 2：“不太小”的班级。设 NVS_i 是个二元变量，如果第 i 个地区学生—教师比是“不太小”的，那么它等于 1。具体来说，若 $STR_i \geq 12$，则 NVS_i 等于 1；否则，NVS_i 等于 0。这个回归也表现出完全多重共线性，但这却是因为一个比前面例子中的回归更微妙的原因。事实上，在我们这个数据集中，没有 $STR_i \leq 12$ 的地区，如在图 4—2 的散点图中可以看到，STR 的最小值是 14。因此，对于所有的观察值，有 $NVS_i = 1$。回想一下，含有截距的线性回归模型可被等价地看做是包含一个对于所有的 i 都等于 1 的回归因子 X_{0i}，如公式(5.6)所示。因此，对于数据集中的所有观察值，我们可以写为 $NVS_i = 1 \times X_{0i}$，也就是说，可以将 NVS_i 写为回归因子的精确的线性组合，具体来说，它等于 X_{0i}。

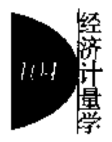
这个例子阐明了多重共线性的两个要点。第一，当回归方程中包含截距时，那么可能涉及完全多重共线性的回归因子之一是“常数”回归因子 X_{0i}。第二，完全多重共线性是对你所掌握的数据集的一种陈述。虽然想象一个少于 12 名学生—教师比的学区是可能的，但是在我们的数据集中不存在这样的学区，所以我们不能在回归中分析它们。

例 3：说英语的人的百分比。设 $PctES_i$ 为第 i 个地区说英语的人的百分比，也可定义为不学习英语的学生的百分比，回归因子仍是完全多重共线的。像前面的例子一样，回归因子之间精确的线性关系涉及“常数”回归因子 X_{0i}：对于每个地区，都有 $PctES_i = 100 \times X_{0i} - PctEL_i$。

这个例子说明了另一点：完全多重共线性是回归因子整体集合的特征。如果从这个回归模型中排除掉截距（即回归因子 X_{0i}）或 $PctEL_i$，那么回归因子不会是完全多重共线的。

完全多重共线性的解决方法。如果设定回归模型时犯了错误，就会出现完全多重共线性。有时这种错误是容易发现的（如在第一个例子中），但有时就不容易发现了（如在第二个例子中）。如果你犯了此类错误，那么软件包会以某种方式让你知道，因为你犯了这样的错误即存在完全多重共线性时，软件是无法计算 OLS 估计量的。

重要概念 5.4



多元回归模型的最小二乘假设

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \cdots, n$$

其中：

1. 给定 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 条件下，u_i 的条件均值为 0，即 $E(u_i | X_{1i}, X_{2i}, \cdots, X_{ki}) = 0$；

2. $(X_{1i}, X_{2i}, \cdots, X_{ki}, Y_i)$，$i=1,\cdots,n$，是取自于它们的联合分布的独立同分布(i. i. d.)的样本；

3. $(X_{1i}, X_{2i}, \cdots, X_{ki}, u_i)$具有非零的有限的四阶矩；

4. 不存在完全多重共线性。

当软件告诉你在你的模型中存在多重共线性时，最为重要的事情是你需要对回归模型进行修正，以消除多重共线性。当存在完全多重共线性时，有些软件是不可靠的。如果回归因子是完全多重共线的，那么至少你可以把对回归因子选择的控制权让给计算机。



不完全多重共线性。尽管名称相似，但不完全多重共线性在概念上与完全多重共线性完全不同。**不完全多重共线性(imperfect multicollinearity)**意味着两个或多个回归因子是高度相关的，在此意义下存在一个与另一个回归因子高度相关的回归因子的线性函数。不完全多重共线性不会给OLS估计量造成任何问题。事实上，当这些回归因子潜在相关时，OLS估计量的目的就是挑选出不同回归因子的独立影响。

多元回归模型的最小二乘假设在重要概念5.4中总结。



5.5 多元回归中OLS估计量的分布

由于数据因样本不同而不同，因此不同的样本会产生不同的OLS估计量的值。对总体回归系数$\beta_0, \beta_1, \cdots, \beta_k$的OLS估计量来说，它们的值在各种可能的样本间变化，便产生了估计量值的不确定性。正如一元回归一样，这种变化被归纳在OLS估计量的抽样分布中。

回想一下4.4节的内容，在最小二乘假设下，一元线性回归模型中的OLS估计量($\hat{\beta}_0$和$\hat{\beta}_1$)是未知系数(β_0和β_1)无偏的、一致的估计量。此外，在大样本条件下，$\hat{\beta}_0$和$\hat{\beta}_1$的抽样分布可被二元正态分布很好地逼近。

这些结论可直接沿用到多元回归分析中。也就是说，在重要概念5.4中的最小二乘假设下，多元线性回归模型中OLS估计量$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$是$\beta_0, \beta_1, \cdots, \beta_k$无偏的、一致的估计量。在大样本条件下，$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$的联合抽样分布被多元正态分布很好地逼近，多元正态分布是从二元正态分布到两个或两个以上的联合正态随机变量分布(见2.4节)的推广。

尽管当存在多个回归因子时，代数运算比较复杂，但是中心极限定理也适用于多元回归模型中的OLS估计量，就像它适用于$\overline{Y}$和适用于存在一个回归因子时OLS估计量的道理一样：OLS估计量$\hat{\beta}_0, \hat{\beta}_1, \cdots, \hat{\beta}_k$是随机样本数据的平均值，如果样本足够大，那些平均值的抽样分布则服从正态分布。由于在数学上多元正态分布用矩阵代数最好处理，因此，OLS估计量联合分布的表达式推迟到第16章介绍。

重要概念5.5总结了这样的结果，即在大样本条件下，多元回归中OLS估计量的分布渐近地服从联合正态分布。一般来讲，OLS估计量之间是相关的，这种相关性是由回归因子之间的相关性引起的。对于有两个回归因子和同方差误差的情况，OLS估计量的联合抽样分布在附录5.2中进行了详细的讨论，而更一般的情况在16.2节中讨论。

OLS估计量的标准误

回想一下，在一元回归的情况下，用样本均值代替期望值来估计OLS估计量的方差是可能的，这就导出了公式(4.19)中给出的估计量$\hat{\sigma}^2_{\hat{\beta}_1}$。在最小二乘假设下，大数定律的结论

隐含着,这些样本均值收敛于其总体均值.例如 $\hat{\sigma}^2_{\hat{\beta}_1}/\sigma^2_{\hat{\beta}_1} \xrightarrow{p} 1$。$\hat{\sigma}^2_{\hat{\beta}_1}$ 的平方根就是 $\hat{\beta}_1$ 的标准误 $SE(\hat{\beta}_1)$,也即 $\hat{\beta}_1$ 的抽样分布标准差的一个估计量。

所有这些都可直接扩展到多元回归。第 j 个回归系数的 OLS 估计量 $\hat{\beta}_j$ 有个标准差,这个标准差是用它的标准误 $SE(\hat{\beta}_j)$ 估计的。标准误的公式使用矩阵最容易说明,所以它在16.2 节中给出。重要的一点是,就标准误来说,单个和多个回归因子的情况之间在概念上不存在差异,最为关键的思想——估计量的大样本正态性和一致地估计其抽样分布标准差的能力——是完全相同的,不论模型中有 1 个、2 个还是有 12 个回归因子。

重要概念 5.5

$\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$ 的大样本分布

如果最小二乘假设(见重要概念 5.4)成立,那么大样本中 OLS 估计量 $\hat{\beta}_0,\hat{\beta}_1,\cdots,\hat{\beta}_k$ 服从联合正态分布,每个 $\hat{\beta}_j$ 服从分布 $N(\beta_j,\sigma_{\hat{\beta}_j}^2)$,$j=0,\cdots,k$。

5.6 单个系数的假设检验和置信区间

本节描述如何检验多元回归方程中关于单个系数的假设和如何构造其置信区间。

5.6.1 单个系数的假设检验

假定保持校区中英语学习者的百分比不变,现在你想要检验"学生—教师比的变化对考试成绩没有影响"这个假设。这相当于假设,在考试成绩对 *STR* 和 *PctEl* 的总体回归中,"学生—教师比的真实系数 β_1 为零"。更一般地说,我们可能想要检验第 j 个回归因子的真实系数 β_j 取某个特定值 $\beta_{j,0}$ 的假设。这个零假设值要么来自于经济理论,要么和学生—教师比的例子一样,来自于应用中的决策背景。如果备择假设是双边的,那么这两个假设在数学上可写成:

$$H_0:\beta_j=\beta_{j,0} \text{ 与 } H_1:\beta_j\neq\beta_{j,0}\text{(双边备择假设)} \tag{5.13}$$

例如,如果第一个回归因子是 *STR*,那么改变学生—教师比对考试成绩没有影响的零假设对应于 $\beta_1=0$ 的零假设(因此 $\beta_{1,0}=0$)。我们的任务就是用样本数据在备择假设 H_1 下检验零假设 H_0。

当存在单个回归因子时,重要概念 4.6 给出了检验这个零假设的程序。该程序的第一步是要计算该系数的标准误。第二步是用重要概念 4.5 中的一般公式计算 t 统计量。第三步是用附表 1 中的累积正态分布计算检验的 p 值,或者比较 t 统计量与对应的检验想要的显著性水平的临界值。这个程序的理论基础是 OLS 估计量有大样本正态分布,在零假设下,其均值为假设真实值,分布的方差可被一致地估计出来。

这个基础也存在于多元回归中。如重要概念 5.5 所述,$\hat{\beta}_j$ 的抽样分布是渐近正态的。在零假设下,此分布的均值为 $\beta_{j,0}$,这个分布的方差能被一致地估计出来。因此,我们可以简单地遵循与单个回归因子情况下相同的程序来检验表达式(5.13)中的零假设。

检验多元回归中单个系数假设的程序,在重要概念 5.6 中总结。在这个重要概念中,实际计算的 t 统计量表示为 t^{act},然而,习惯上将它简记为 t,本书以下部分中,我们采用这个简化的符号。

5.6.2 单个系数的置信区间

多元回归模型中构造置信区间的方法也和单个回归因子模型中的一样。这个方法已在重要概念 5.7 中总结。

重要概念 5.6 中构造假设检验的方法和重要概念 5.7 中构造置信区间的方法,依赖于 OLS 估计量 $\hat{\beta}_j$ 分布的大样本正态近似。因此,应当牢记,这些量化抽样不确定性的方法只有在大样本条件下才能确保有效。

重要概念 5.6

检验假设 $\beta_j=\beta_{j,0}$ 和备择假设 $\beta_j\neq\beta_{j,0}$

1. 计算 $\hat{\beta}_j$ 的标准误 $SE(\hat{\beta}_j)$。
2. 计算 t 统计量:

$$t=\frac{\hat{\beta}_j-\beta_{j,0}}{SE(\hat{\beta}_j)} \tag{5.14}$$

3. 计算 p 值:

$$p\text{ 值}=2\Phi(-|t^{act}|) \tag{5.15}$$

其中,t^{act} 是实际计算的 t 统计量的值。如果 p 值小于 0.05,或者说,如果 $|t^{act}|>1.96$,那么在 5% 的显著水平下拒绝零假设。

检验 $\beta_j=0$ 的标准误,特别是 t 统计量和 p 值,可由回归软件自动计算出来。

5.6.3 在考试成绩和学生—教师比案例中的应用

一旦我们控制了地区中英语学习者的百分比,我们能拒绝学生—教师比的变化对考试成绩没有影响的零假设吗?控制英语学习者的百分比后,学生—教师比的变化对考试成绩影响的 95% 的置信区间是什么?我们现在能够得出结论。由 OLS 估计出的考试成绩对 *STR* 和 *PctEL* 的回归方程在公式(5.10)中给出,这里我们把它重写出来,标准误在系数下面的括号里给出。

$$\widehat{TestScore}=\underset{(8.7)}{686.0}-\underset{(0.43)}{1.10}\times STR-\underset{(0.031)}{0.650}\times PctEL \tag{5.16}$$

为了检验 *STR* 的真实系数为 0 的假设,我们首先需要计算公式(5.14)中的 t 统计量。因为零假设假定这个系数的真实值为 0,所以 t 统计量是 $t=(-1.10-0)/0.43=-2.54$。相应地,p 值为 $2\Phi(-2.54)=1.1\%$,也就是说,我们能够拒绝零假设的最小显著性水平为 1.1%。由于 p 值小于 5%,因此在 5% 的显著性水平下能够拒绝零假设(但在 1% 的显著性水平下却不一定能拒绝)。

STR 总体系数的 95% 的置信区间为 $-1.10\pm1.96\times0.43=(-1.95,-0.26)$,也就是说,我们有 95% 的把握相信系数的真实值介于 -1.95 和 -0.26 之间。对教育主管所感兴趣的将学生—教师比减少 2 个单位这一问题而言,这个减少对考试成绩影响的 95% 的置信区间为 $(-1.95\times2,-0.26\times2)=(-3.90,-0.52)$。

重要概念 5.7

多元回归中单个系数的置信区间

系数 β_j 的 95% 的双边置信区间,是指在 95% 的概率保证下,包含 β_j 的真实值的区间,



也就是说，在所有可能随机抽取的样本中，有95%的样本β_j的真实值落于其中，或者说，它也是不能被5%的双边假设检验所拒绝的β_j的值的集合。当样本容量很大时，95%的置信区间是：

$$\beta_j \text{ 的95\%的置信区间} = (\hat{\beta}_j - 1.96SE(\hat{\beta}_j), \hat{\beta}_j + 1.96SE(\hat{\beta}_j)) \quad (5.17)$$

用1.645代替公式(5.17)中的1.96就可以得到90%的置信区间。

把每个学生的费用加入到方程中　根据公式(5.16)中的多元回归分析结果，已经可以说服该教育主管，即根据当前所有的证据，在其所属的地区减小班级规模将会对提高考试成绩有所帮助。但是，现在她转向了一个更加细微的问题。如果她想要雇佣更多的教师，她可以通过削减预算的其他方面（如不配置新的电脑、减少维护费用等），或者要求增加她的财政预算来支付那些教师的工资，而纳税人不喜欢这样。她会问，在保持每个学生的费用（和英语学习者的百分比）不变的情况下，减少学生—教师比对考试成绩的影响是什么？

这个问题可以通过估计考试成绩对学生—教师比、每个学生的总费用和英语学习者的百分比的回归来解决。OLS回归线为：

$$\widehat{TestScore} = \underset{(15.5)}{649.6} - \underset{(0.48)}{0.29} \times STR + \underset{(1.59)}{3.87} \times Expn - \underset{(0.032)}{0.656} \times PctEL \quad (5.18)$$

其中，*Expn*是以千美元表示的地区中每个学生的年度总费用。

结果是惊人的。在保持每个学生的费用和英语学习者的百分比不变的情况下，改变学生—教师比对考试成绩的影响非常小：公式(5.16)中*STR*的估计系数为-1.10，但是在公式(5.18)中增加了*Expn*作为回归因子后，这个估计系数仅为-0.29。此外，检验该系数的真实值为零的t统计量现在是$t = (-0.29 - 0)/0.48 = -0.60$，所以，这个系数的总体值确实为零的假设不能被拒绝，即使在10%的显著性水平下($|-0.60| < 1.645$)也不能被拒绝。因此，如果每个学生的总费用保持不变，那么，公式(5.18)没有为雇佣更多的教师会提高考试成绩提供任何证据。

注意，当*Expn*变量被添加进来后，*STR*的标准误从公式(5.16)中的0.43增加到公式(5.18)中的0.48。这说明了一个要点，即回归因子之间的相关性（*STR*和*Expn*之间的相关系数为-0.62）会使OLS估计量变得更不精确（进一步讨论见附录5.2）。

我们那些愤怒的纳税人会怎么样呢？他们断言学生—教师比系数(β_1)和每个学生的费用系数(β_2)的总体值都是零，也就是说，他们假设$\beta_1 = 0$和$\beta_2 = 0$。尽管看起来我们好像能够拒绝这个假设，因为公式(5.18)中检验$\beta_2 = 0$的t统计量为$t = 3.87/1.59 = 2.43$，但是这个推理有缺陷。纳税人的假设是一个联合假设，要检验这个假设，我们需要一个新的工具——*F*统计量。

5.7 联合假设的检验

本节介绍如何对多元回归系数做联合假设，以及如何用*F*统计量对它们进行检验。

5.7.1 检验两个或多个系数的假设

联合零假设　考虑公式(5.18)中考试成绩对学生—教师比、每个学生的费用和英语学习者百分比这三个变量的回归。我们愤怒的纳税人假设，一旦我们控制了英语学习者的百分比，学生—教师比和每个学生的费用对考试成绩都没有影响。由于*STR*是公式(5.18)中

5.7.2 F统计量

F统计量(F statistic)用来检验关于回归系数的联合假设。F统计量的公式已被嵌入到现代回归分析软件中。我们首先讨论两个约束条件的情形,然后再转向q个约束条件的一般情形。

含有$q=2$个约束条件的F统计量。当联合零假设含有两个约束条件$\beta_1=0$和$\beta_2=0$时,F统计量用下面的公式将t统计量t_1和t_2结合在一起:

$$F=\frac{1}{2}\left(\frac{t_1^2+t_2^2-2\hat{\rho}_{t_1,t_2}t_1t_2}{1-\hat{\rho}_{t_1,t_2}^2}\right) \tag{5.21}$$

其中,$\hat{\rho}_{t_1,t_2}$是两个t统计量之间相关系数的估计量。

为了理解公式(5.21)中的F统计量,首先假设我们知道t统计量是不相关的,因此我们就可以丢掉与$\hat{\rho}_{t_1,t_2}$相关的项。如果是这样的话,公式(5.21)就可简化为$F=\frac{1}{2}(t_1^2+t_2^2)$,即$F$统计量是$t$统计量的平方和的平均值。在零假设下,$t_1$和$t_2$是相互独立的标准正态随机变量(根据假设,$t$统计量是不相关的),所以在零假设下,$F$服从$F_{2,\infty}$分布(见2.4节)。在备择假设要么$\beta_1$不为零,要么$\beta_2$不为零(或者两者都不为零)的条件下,$t_1^2$或$t_2^2$(或二者)将会很大,这将会导致检验结果是拒绝零假设。

一般来讲,t统计量是相关的,公式(5.21)中F统计量的公式就是对这个相关性进行了调整。进行这样调整的目的是,在零假设下,不论t统计量是否相关,F统计量在大样本条件下都服从$F_{2,\infty}$分布。

含有q个约束条件的F统计量。检验表达式(5.20)中q个约束条件的联合零假设的F统计量的公式在16.3节中给出。这个公式已被嵌入到回归分析软件中,使得F统计量在实际中易于计算。

在零假设下,F统计量具有抽样分布,其抽样分布在大样本条件下是由$F_{q,\infty}$的分布给出的。也就是说,在大样本中,在零假设下有:

$$F\text{统计量服从分布}\ F_{q,\infty} \tag{5.22}$$

因此,对于适当的q值和想要的显著性水平,F统计量的临界值可以从附表4中的$F_{q,\infty}$分布中得到。

利用F统计量计算p值。F统计量的p值,可以使用其分布的大样本χ^2近似计算出来。设F^{act}为实际计算的F统计量的值,由于在零假设下F统计量服从大样本的$F_{q,\infty}$分布,因此p值为:

$$p\text{值}=\Pr(F_{q,\infty}>F^{act}) \tag{5.23}$$

公式(5.23)中的p值可以使用$F_{q,\infty}$分布表(或χ_q^2分布表,因为χ_q^2分布的随机变量是$F_{q,\infty}$分布的随机变量的q倍)来求值。另一种可供选择的方法是,使用计算机来求p值,因为累积卡方分布和F分布的公式已被嵌入到最新的统计软件中了。

"全体"回归的F统计量。"全体"回归的F统计量检验所有斜率系数都为零的联合假设,即零假设和备择假设是:

$$H_0:\beta_1=0,\beta_2=0,\cdots,\beta_k=0\ \text{与}\ H_1:\beta_j\neq 0,\text{至少有一个}\ j(j=1,\cdots,k) \tag{5.24}$$

在这个零假设下,没有一个回归因子解释了Y_i中的任何变差,尽管截距(在零假设下它是Y_i的均值)可能是非零的。表达式(5.24)中的零假设是表达式(5.20)中一般零假设的特殊情形,且全体回归的F统计量是在表达式(5.24)中对零假设计算的F统计量。在大样

本条件下，全体回归的 F 统计量服从 $F_{q,\infty}$ 的分布。

当 $q=1$ 时的 F 统计量　当 $q=1$ 时，F 统计量检验一个约束条件。于是，联合零假设变成了单个回归系数的零假设，F 统计量是 t 统计量的平方。

异方差和同方差　回想一下 4.9 节，由于历史原因，统计软件有时将仅适用于同方差的标准误作为默认值计算，因此，要计算异方差稳健的标准误，用户必须自己设定。这个提醒也适用于 F 统计量：为了确保你使用异方差稳健的 F 统计量，在有些回归软件包中，你必须选择“稳健的”选项，使用“协方差阵”的稳健估计值。如果使用 F 统计量的仅适用于同方差的形式（在附录 5.3 中讨论），但误差是异方差的，那么在零假设下，F 统计量不会服从表达式(5.22)中的 $F_{q,\infty}$ 分布，这会导致令人误解的统计推断。

5.7.3　在考试成绩和学生—教师比案例中的应用

我们现在可以检验下面的零假设和备择假设。零假设是：学生—教师比的系数和平均每个学生的费用系数均为零。备择假设是：它们之中至少有一个不为零。同时要控制该地区内英语学习者的百分比。

为了检验这个假设，我们需要用公式(5.18)中的考试成绩对 *STR*、*Expn*、*PctEL* 的回归来计算检验 $\beta_1=0$ 和 $\beta_2=0$ 的 F 统计量。这个 F 统计量为 5.43。在零假设下，这个统计量在大样本条件下服从 $F_{2,\infty}$ 分布。$F_{2,\infty}$ 分布的 5% 的临界值为 3.00（见附表 4），1% 的临界值为 4.61。根据数据所计算的 F 统计量的值为 5.43，超过了 4.61，所以零假设在 1% 的显著性水平下被拒绝。如果零假设确实是真的，那么我们抽取一个会得出和 5.43 一样大的 F 统计量的样本是非常不可能的（p 值为 0.005）。根据公式(5.18)中的证据以及这个 F 统计量所总结的信息，在保持英语学习者的百分比不变的条件下，我们可以拒绝纳税人提出的“学生—教师比及平均每个学生的费用对考试成绩都没有影响”的假设。

5.8　检验涉及多个系数的单个约束条件

有时经济理论会提出涉及两个或多个回归系数的单个约束条件。例如，理论可能会提出一个 $\beta_1=\beta_2$ 形式的零假设，即第一个和第二个回归因子的影响是相同的。在这种情况下，我们的任务就是检验两个系数相同的零假设和两个系数不同的备择假设，即：

$$H_0:\beta_1=\beta_2 \text{ 与 } H_1:\beta_1\neq\beta_2 \tag{5.25}$$

这个零假设有一个约束条件，所以 $q=1$，但这个约束条件涉及多个系数（β_1 和 β_2）。我们需要修正现存的方法来检验这个假设。有两种方法，哪一种方法最容易取决于你所用的软件。

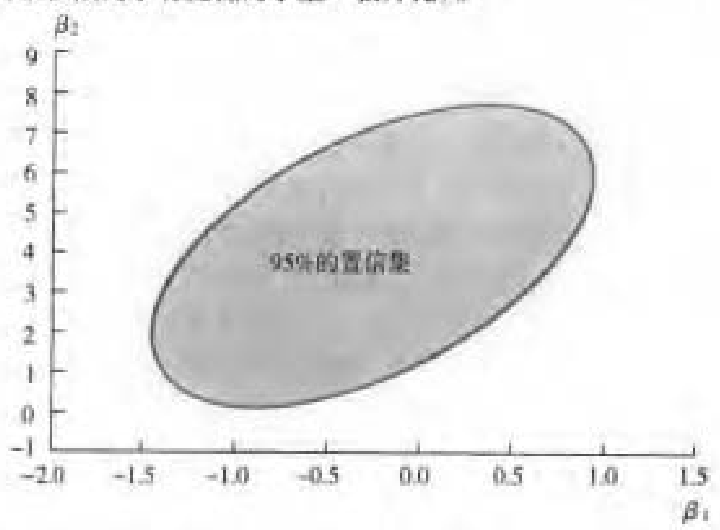
方法 1：直接检验约束。某些统计软件包含有某个专门的设计命令来检验像表达式(5.25)那样的约束条件，其结果是一个 F 统计量，由于 $q=1$，所以在零假设下该统计量服从 $F_{1,\infty}$ 分布。（回想一下 2.4 节，一个标准正态随机变量的平方服从 $F_{1,\infty}$ 分布，所以 $F_{1,\infty}$ 分布的 95% 的百分位数是 $1.96^2=3.84$。）



方法 2：变换回归形式。如果你的统计软件包不能直接检验约束条件，那么可以将原来的回归方程重写一下，以把表达式(5.25)中的约束条件变换为一个单个回归系数的约束条件，用这样的技巧来检验表达式(5.25)中的假设。具体来说，假设回归中只有两个回归因子 X_{1i} 和 X_{2i}，所以总体回归方程的形式为：

$$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i \tag{5.26}$$

英语学习者的百分比不变的情况下，学生一教师比和平均每个学生费用系数的一个95%的置信集（置信椭圆）。这个椭圆不包括(0,0)点，这意味着利用F统计量，这两个系数都为零的零假设在5%的显著性水平下被拒绝了，这一点在5.7节中我们就已经知道了。置信椭圆像个胖的香肠，香肠的两端分别指向左下方和右上方。这种指向的原因是，所估计的$\hat{\beta}_1$和$\hat{\beta}_2$之间的相关系数是正的，它是由回归因子STR和$Expn$之间的负相关所引起的（每个学生的花费越多的学校倾向于有更低的学生一教师比）。



注：β_1和β_2的95%的置信集是个椭圆。这个椭圆包含用F统计量在5%的显著性水平下不能拒绝的β_1和β_2的数值对。

图5—1 β_1和β_2的95%的置信集

5.10 其他一些回归统计量

多元回归中三个常用的描述性统计量是：回归的标准误、回归的R^2和调整的R^2（也称为$\bar{R}^2$）。这三个统计量都测度了多元回归线的OLS估计值描述或"拟合"数据的好坏程度。

5.10.1 回归的标准误（SER）

回归的标准误估计了误差项u_i的标准差。因此，SER是Y围绕其回归线分布的离散程度的一个测度。在多元回归中，SER是：

$$SER = s_{\hat{u}}，其中\ s_{\hat{u}}^2 = \frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2 = \frac{SSR}{n-k-1} \tag{5.28}$$

这里，SSR是残差平方和，$SSR = \sum_{i=1}^{n}\hat{u}_i^2$。

表达式(5.28)中的定义和4.8节中单个回归因子模型中SER的定义之间的惟一差异是，这里的除数是$n-k-1$而不是$n-2$。在4.8节中，除数$n-2$（不是n）调整了由估计两个系数（回归线的斜率和截距）所引入的向下偏差。这里，除数$n-k-1$调整了由估计$k+1$个系数（k个斜率系数加上截距）所引入的向下偏差。与4.8节中一样，使用$n-k-1$而不使用n被称为自由度调整。如果只有单个回归因子，那么$k=1$，因此4.8节中的公式和表达式(5.28)中的公式相同。当n很大时，自由度调整的影响是可以忽略的。

5.10.2 R^2

回归的 R^2 是被回归因子所解释(或所预测)的 Y_i 的样本方差的部分,或者说,R^2 是1减去没有被回归因子所解释的 Y_i 的方差的部分。

R^2 的数学定义和一元回归中的定义一样:

$$R^2 = \frac{ESS}{TSS} = 1 - \frac{SSR}{TSS} \tag{5.29}$$

这里,被解释平方和为 $ESS = \sum(\hat{Y}_i - \bar{Y})^2$,总平方和为 $TSS = \sum(Y_i - \bar{Y})^2$。

在多元回归中,只要增加一个回归因子,R^2 的值就会增加,除非这个新的回归因子与原来的回归因子之间是完全多重共线的。为了弄清楚这一点,首先考虑采用一个回归因子,然后再增加另一个回归因子。当你用 OLS 估计含有两个回归因子的模型时,OLS 可以解出使残差平方和最小的系数值。如果 OLS 求出那个新的回归因子的系数恰好为零,那么不论回归中是否包含第二个变量,SSR 的值都将是相同的。但如果 OLS 求出的是任何非零的值,那么相对于不包括这个回归因子的回归来说,一定是这个值使 SSR 降低了。在实际中,估计系数恰好为零的情况是极其少见的,因此一般而言,当增加一个新的回归因子时,SSR 将会减小。另一方面,这意味着当增加一个新的回归因子时,R^2 通常会增加(但从不会减少)。

5.10.3 “调整的 R^2”

由于当一个新变量被添加进来时 R^2 会增加,因此,R^2 的增加并不意味着增加一个变量确实会改进模型的拟合程度。在这个意义下,R^2 给出了一个回归对数据的拟合程度的夸大估计。纠正它的一个方法是用某个因子来缩减或降低 R^2,这就是调整的 R^2 或 $\bar{R}^2$ 的含义。

调整的 R^2(adjusted R^2)或 $\bar{R}^2$ 是 R^2 的一个修正形式。当增加一个新的回归因子时,它不一定会增加。$\bar{R}^2$ 的计算公式是:

$$\bar{R}^2 = 1 - \frac{n-1}{n-k-1} \cdot \frac{SSR}{TSS} = 1 - \frac{s_{\hat{u}}^2}{s_Y^2} \tag{5.30}$$

这个公式与公式(5.29)中 R^2 的第二个定义的不同之处是,残差平方和与总平方和之比,乘上了因子 $(n-1)/(n-k-1)$。如公式(5.30)中的第二个表达式所示,这使得调整的 R^2 等于1减去 OLS 残差的样本方差(含有表达式(5.28)中的自由度修正)与 Y 的样本方差之比。

关于 $\bar{R}^2$,有三件事情需要知道:

第一,$(n-1)/(n-k-1)$ 总是大于1,所以 $\bar{R}^2$ 总是小于 R^2。

第二,增加一个回归因子对 $\bar{R}^2$ 有两个互为相反的影响。一方面,SSR 下降会使 $\bar{R}^2$ 增加;另一方面,因子 $(n-1)/(n-k-1)$ 会增加。此时,$\bar{R}^2$ 是增加还是减少取决于这两个影响中哪一个更强。

第三,$\bar{R}^2$ 可能是负的。综合考虑,当回归因子使残差平方和减少很小的部分,以至于这个减少的量无法弥补因子 $(n-1)/(n-k-1)$ 的变化时,就会发生这种情况。

5.10.4 R^2 和调整后的 R^2 的实际意义

R^2 或 $\bar{R}^2$ 接近于1,意味着回归因子预测样本中因变量值的能力很强;若 R^2 或 $\bar{R}^2$ 接近于0,则意味着它们不能有效地预测因变量的变化。这使得这两个统计量成为评价回归预测能力的有效测度。但是,它们的含义远非看上去那么简单。

使用 R^2 或 $\bar{R}^2$ 时,要当心四个陷阱:

1. R^2 或 $\bar{R}^2$ 的增加并不一定意味着增加的变量在统计上是显著的。不管它在统计上是否是显著的，只要我们增加一个回归因子，R^2 的值都会增加。$\bar{R}^2$ 的值不会总是增加，但如果它确实增加了，那么这并不一定意味着增加的回归因子的系数在统计上是显著的。要确认一个增加的变量在统计上是否是显著的，你需要用 t 统计量进行假设检验。

2. 高的 R^2 或 $\bar{R}^2$ 并不意味着回归因子是因变量的真实原因。想象一下，我们曾用考试成绩对每个学生的停车场面积进行回归。停车场的面积和学生—教师比相关，和郊区或市区的学校相关，也可能和地区的收入水平相关，许多事情都可能和考试成绩相关。因此，考试成绩对每个学生停车场面积的回归，可能有很高的 R^2 或 $\bar{R}^2$，但是这个关系不是因果关系。（想办法告诉教育主管，提高考试成绩的方法就是增加停车空间！）

3. 高的 R^2 或 $\bar{R}^2$ 并不意味着不存在遗漏变量偏差。回想一下 5.1 节中的讨论，这些讨论主要涉及考试成绩对学生—教师比回归中的遗漏变量偏差问题。这个回归的 R^2 的作用永远不会上升，因为它在这个讨论中不起什么逻辑作用。遗漏变量偏差在低 R^2、适度 R^2 或高 R^2 的回归中都可能发生。反之，低的 R^2 并不意味着一定存在遗漏变量偏差。

4. 高的 R^2 或 $\bar{R}^2$ 并不一定意味着你抓到了最合适的回归因子集，低的 R^2 或 $\bar{R}^2$ 也不一定意味着你选择的回归因子集不合适。在多元回归中，什么样的回归因子集是最恰当的？这是一个很难回答的问题，我们讲完整本书之后再回到这个问题。回归因子的确定必须权衡遗漏变量偏差、数据可获得性和数据质量，而最重要的是经济理论和所要解决的问题的性质。这些问题中没有一个能用高（或低）的回归 R^2 或 $\bar{R}^2$ 来简单地回答。

这些要点在重要概念 5.8 中总结。

5.11 遗漏变量偏差与多元回归

如果一个遗漏的 Y_i 的决定性因素至少和一个回归因子相关，那么多元回归系数的 OLS 估计量就会存在遗漏变量偏差。例如，来自富裕家庭的学生通常比那些不太富裕的同学有更多的学习机会，这也可能导致他们有好的考试成绩。此外，如果该地区比较富裕，那么学校倾向于有较大的预算和较低的学生—教师比。如果是这样的话，那么学生的富裕程度和学生—教师比将会是负相关的。即使在控制了英语学习者的百分比后，学生—教师比系数的 OLS 估计值也将会提高地区平均收入对考试成绩的影响。简而言之，遗漏学生的经济背景可能会使考试成绩对学生—教师比和英语学习者的百分比的回归中存在遗漏变量偏差。

多元回归中遗漏变量偏差的一般条件和一元回归中的条件类似：如果遗漏变量是 Y_i 的一个决定性因素，如果它至少与一个回归因子相关，那么 OLS 估计量将会存在遗漏变量偏差。就像在 5.6 节中所讨论的，OLS 估计量是相关的，因此，一般而言，所有系数的 OLS 估计量将会是有偏的。多元回归中关于遗漏变量偏差的两个条件在重要概念 5.9 中总结。

在数学上，如果满足遗漏变量偏差的两个条件，那么至少有一个回归因子与误差项相关。这意味着给定 $X_{1i},\cdots,X_{ki}$ 条件下 u_i 的条件期望值是非零的，这样就违背了第一个最小二乘假设。因此，即使样本容量很大，遗漏变量偏差仍继续存在，也就是说，遗漏变量偏差隐含着 OLS 估计量是不一致的。



重要概念 5.8

R^2 和 $\bar{R}^2$：它们告诉了你什么？它们又没告诉你什么

R^2 和 $\bar{R}^2$ 告诉你回归因子是否有效地解释了或预测了所研究的样本数据中因变量的值。如果R^2（或$\bar{R}^2$）接近于1，那么在这个样本中回归因子会很好地预测因变量的变化，其

许多因素潜在地影响着一个地区的平均考试成绩。一些影响考试成绩的因素可能与学生—教师比相关,所以回归中遗漏它们将会导致遗漏变量偏差。如果可以获得这些遗漏变量的数据,那么解决这个问题的方法就是把它们作为额外的回归因子包含在多元回归中。当我们这样做时,学生—教师比的系数,就是在保持这些其他因素不变的条件下,反映学生—教师比的变化对考试成绩的影响。

这里,我们考虑三个可能影响考试成绩的控制学生背景特征的变量。这些控制变量之一是我们在前面已用过的一个变量,即仍在学习英语的学生的比重。其他两个变量是新的,用来控制学生的经济背景。在数据集中,没有对学生经济背景进行描述的较完善的指标,因此,我们引入用来反映地区低收入状况的两个不很完善的指标。第一个新的变量是在学校里有资格得到补助或免费午餐计划的学生的百分比。如果学生的家庭收入低于一个特定的收入限值(大约为贫穷线的150%),那么他们就有资格享受这个计划。第二个新的变量是地区中学生家庭有资格享受加利福尼亚州收入援助计划的学生百分比。家庭是否有资格享受这项收入援助计划部分地取决于他们的家庭收入,但是其收入限值比午餐补助计划的收入限值低(更严格)。因此,这两个变量测度了该地区经济条件不好的学生的比例,虽然它们是相关的,但是它们不是完全相关的(它们的相关系数是0.74)。尽管理论表明经济背景可能是一个重要的遗漏因素,但是理论和专家判断实际上没有帮我们决定这两个变量中(有资格享有补助午餐的学生百分比或有资格享有收入补助的学生百分比)哪一个是经济背景的较好测度。就我们的基准设定而言,我们选择有资格享有午餐补助计划的学生百分比作为经济背景变量,但我们也考虑了包含其他变量的备择设定。

考试成绩和这些变量的散点图在图5—2中给出。这些变量中每一个都显示出与考试成绩的负相关关系。考试成绩和英语学习者百分比之间的相关系数为-0.64;考试成绩和有资格享有午餐补助计划的学生百分比之间的相关系数为-0.87;考试成绩和有资格享有收入援助计划的学生百分比之间的相关系数为-0.63。

现在我们面临一个交流的问题。什么样的方式是表现包含多个回归因子数据子集的多个多元回归方程结果的最好方式?到目前为止,我们是用写出像公式(5.18)那样所估计的回归方程的形式提供回归结果。当只有几个回归因子和几个回归方程时,这样做是可以的,但是在有多个回归因子和多个方程的情况下,这种表示方法可能会不好交流,也不好理解。看来,交流多个回归方程结果的一个比较好的方法是用表格。

表5—2中归纳了考试成绩对不同子集的回归因子回归的结果。每一列总结了一个单独的回归,每个回归都有相同的因变量——考试成绩。前五行中的子项是所估计的回归系数,它们的标准误在其下面的括号里给出。星号表示检验相关的系数为零的假设,其t统计量在5%的水平下是否是显著的(*),或在1%的水平下是否是显著的(**)。表5—2的最后三行包含了回归的总括性统计量(回归的标准误、SET和调整的R^2即$\bar{R}^2$)和样本容量(它对于所有的回归都是一样的,即420个观察值)。

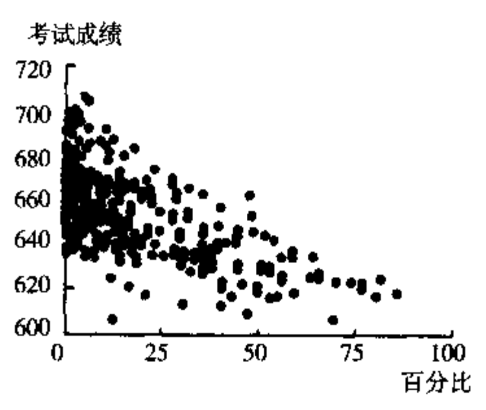
到目前为止,我们以方程形式给出的所有信息都列在这个表格的一列中。例如,考虑没有控制变量时的考试成绩对学生—教师比的回归。用方程形式表示,这个回归是:

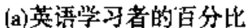
$$\widehat{TestScore} = \underset{(10.4)}{698.9} - \underset{(0.52)}{2.28} \times STR, \bar{R}^2 = 0.049, SER = 19.26, n = 420 \quad (5.31)$$



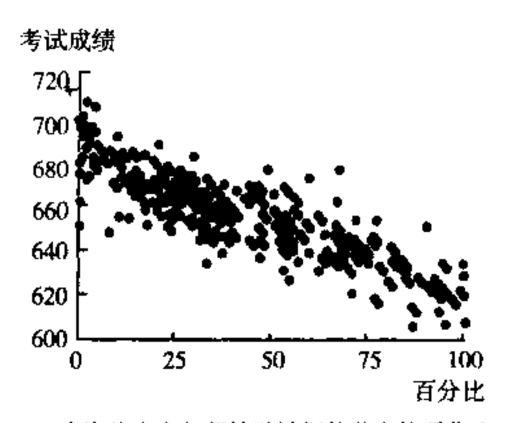
所有这些信息都显示在表5—2的第(1)列中。学生—教师比的估计系数(-2.28)出现在第一行的子项中,而它的标准误(0.52)就出现在估计系数下面的括号里。截距(689.9)和它的标准误(10.4)在标有“截距”的那一行中给出(有时你会看到这一行标有“常数”,因为



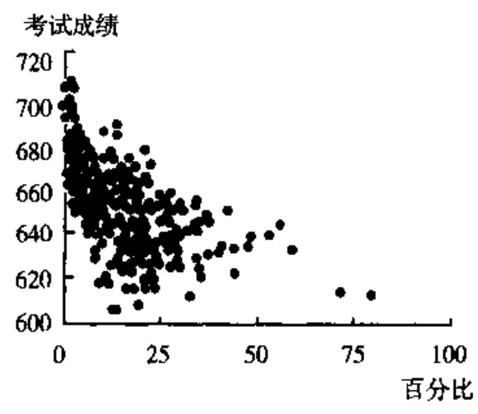




(a)英语学习者的百分比



(b)有资格享有午餐补助计划的学生的百分比



(c)有资格享有收入援助计划的学生的百分比

注:散点图显示了考试成绩和下列指标间负相关的关系。即:(a)与英语学习者的百分比(相关系数 = -0.64);(b)与有资格享有午餐补助计划的学生的百分比(相关系数 = -0.87);(c)与有资格享有收入援助计划的学生的百分比(相关系数 = -0.63)。

图5—2 考试成绩对三种学生特征的散点图

如5.2节中所讨论的,截距可被看做为总是等于1的回归因子的系数)。同理,$\bar{R}^2$(0.049),SER(18.58)和样本容量n(420)在最后一行中出现。其他回归因子行中空白的子项表示,那些回归因子不包含在这个回归中。

尽管表5—2没有报告t统计量,但是这些能够根据所提供的信息计算得到,例如,检验第(1)列中学生—教师比的系数为零的假设的t统计量是$(-2.28)\div 0.52=-4.38$。这个假设在1%的显著性水平下被拒绝,它用表中估计系数后面的"***"来表示。

包含测度学生特征的控制变量的回归报告在第(2)~第(5)列中给出。如前面公式(5.16)所述,第(2)列报告了考试成绩对学生—教师比和英语学习者的百分比的回归。

第(3)列提供了基准设定,其中回归因子是学生—教师比和两个控制变量——英语学习者的百分比与享有免费午餐的学生的百分比。

第(4)列和第(5)列提供了备择设定,它研究了已被测度的学生经济背景情况的变化对考试成绩的影响。在第(4)列中,得到公众援助的学生的百分比作为一个回归因子被包含进来,而在第(5)列中,两个经济背景变量都被包含进来了。

这些结果给出了三个结论:

1. 控制这些学生特征之后,学生—教师比对考试成绩的影响大约减少了一半。这个估计的影响对哪一个特定的控制变量被包含在回归中不是非常敏感的。在所有的情形下,学



生—教师比的系数在5%的水平下在统计上都是显著的。在含有控制变量的四个设定的回归(2)~(5)列中,如果保持学生特征不变,对学生—教师比这一指标来说,平均每名教师减少1名学生,那么估计的结果是使平均考试成绩提高大约1分。

表5—2　　考试成绩对学生—教师比和学生特征控制变量的回归结果

(利用加利福尼亚州小学学区的数据)

因变量:学区的平均考试成绩					
回归因子	(1)	(2)	(3)	(4)	(5)
学生—教师比(X_1)	−2.28**	−1.10*	−1.00**	−1.31**	−1.01**
	(0.52)	(0.43)	(0.27)	(0.34)	(0.27)
英语学习者的百分比(X_2)	−0.650**	−0.122**	−0.488**	−0.130**	
	(0.031)	(0.033)	(0.030)	(0.036)	
享有午餐补助计划的学生的百分比(X_3)			−0.547**		−0.529**
			(0.024)		(0.038)
享有公众收入援助计划的学生的百分比(X_4)				−0.790**	0.048
				(0.068)	(0.059)
截距	698.9**	686.0**	700.2**	698.0**	700.4**
	(10.4)	(8.7)	(5.6)	(6.9)	(5.5)
概括性统计量					
SER	18.58	14.46	9.08	11.65	9.08
$\bar{R}_2$	0.049	0.424	0.773	0.626	0.773
n	420	420	420	420	420

注:这些回归是用附录4.1中所介绍的加利福尼亚州的$K-8$学区的数据来估计的。标准误在系数下面的括号中给出。利用双边检验,单个系数在*5%的显著性水平或**1%的显著性水平下在统计上是显著的。

2. 学生特征的变量对于考试成绩是非常有用的预测因子。学生—教师比仅仅解释了考试成绩变化的一小部分:第(1)列中的$\bar{R}^2$为0.049。然而,当增加学生特征变量时,$\bar{R}^2$会跳跃式地增加。例如,基准设定回归(3)中的$\bar{R}^2$值为0.773。学生人口统计变量的系数的符号与图5—2中的样式是一致的:有很多英语学习者的地区和有许多穷孩子的地区有较低的考试成绩。

3. 个别地看,控制变量在统计上不总是显著的。在设定(5)中,享有收入援助的学生的百分比的系数为零的假设在5%的水平下(t统计量是−0.82)没有被拒绝。因为在基准设定(3)的回归中,增加这个控制变量对估计系数及其标准误的影响是可以忽略的,而且因为在设定(5)中这个控制变量的系数在统计上不是显著的,所以至少对这个分析的目的而言,这个新增的控制变量是多余的。

5.13 结论

本章以一个令人忧虑的问题开始:在考试成绩对学生—教师比的回归中,被遗漏的那些影响考试成绩的学生特征可能和地区中的学生—教师比相关,而且如果是这样的话,这些地区中的学生—教师比会增加这些遗漏的学生特征对考试成绩的影响。因此,OLS估计量将会存在遗漏变量偏差。为了缓解这个潜在的遗漏变量偏差,我们通过将控制不同学生特征的变量包含在回归中(英语学习者的百分比和学生经济背景的两个测度)来扩展回归模型。这样做会使学生—教师比单位变化的估计影响减少一半,尽管在保持这些控制变量不变的情况下,它仍有可能在5%的显著性水平下拒绝对考试成绩的总体影响为零的零假设。因为它们消除了由这些学生特征所引起的遗漏变量偏差,所以就为教育主管提供建议而言,这

些多元回归估计值(及相关的置信区间)比第4章中的单个回归因子的估计值要有用得多。

本章中的分析已假设回归因子中的总体回归函数是线性的,也就是说,给定回归因子,Y_i 的条件期望是一条直线。但是,这样考虑并没有什么特定的原因。实际上,在大班地区减少学生—教师比的影响可能非常不同于对小班地区的影响。如果是这样的话,总体回归线与变量 X 之间就不是线性的关系,而是关于变量 X 的非线性函数。然而,要将我们的分析推广到变量 X 的非线性回归函数情形,我们需要应用下一章所开发的工具。

总结

1. 遗漏变量偏差发生的条件是:(1)与引人的回归因子相关;(2)是 Y 的一个决定性因素。

2. 多元回归模型是包含多个回归因子 $X_1,\cdots,X_k$ 的线性回归模型。与每个回归因子相联系的是回归系数 $\beta_1,\cdots,\beta_k$。系数 β_1 的含义是,在保持其他回归因子不变的情况下,与 X_1 的单位变化相联系的 Y 的期望变化。其他回归系数的解释与此类似。

3. 多元回归中的系数可以用 OLS 进行估计。当满足重要概念5.4中的四个最小二乘假设时,OLS 估计量在大样本条件下是无偏的、一致的且服从正态分布。

4. 单个回归系数的假设检验和置信区间的构建,可以用和第4章的单变量线性回归模型中所使用的本质上相同的程序来进行。例如,β_1 的95%的置信区间由 $\hat{\beta}_1+1.96SE(\hat{\beta}_1)$ 给出。

5. 涉及一个以上的系数约束条件的假设被称为联合假设。联合假设可以用 F 统计量进行检验。

6. 回归的标准误 R^2 和 $\bar{R}^2$ 是多元回归模型的总括性描述统计量。

重要术语

遗漏变量偏差　多元回归模型　总体回归线　总体回归函数　截距　X_{1i}的系数　控制变量　偏效应　总体多元回归模型　同方差　异方差　$\beta_1,\cdots,\beta_k$ 的 OLS 估计量　OLS 回归线　预测值　OLS 残差　完全多重共线性　不完全多重共线性　约束条件　联合假设　F 统计量　95%的置信集　R^2 和调整的 $R^2(\bar{R}^2)$　基准设定　备择设定　经验规则 F 统计量

复习概念

5.1　研究人员对计算机的使用情况对考试成绩的影响感兴趣。利用本章中所用的校区数据,该研究人员用地区的平均考试成绩对平均每个学生的计算机数进行回归。在分析增加学生人均计算机数对考试成绩的影响时,估计量 $\hat{\beta}_1$ 是无偏估计量吗?为什么是或为什么不是?如果你认为 $\hat{\beta}_1$ 是有偏的,那么它是向上偏还是向下偏?为什么?



5.2　一个多元回归包含两个回归因子:$Y_i=\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+u_i$。如果保持 X_2 不变,X_1 增加3个单位对 Y 的期望变化是多少?如果保持 X_1 不变,X_2 减少5个单位对 Y 的期望变化是多少?如果 X_1 增加3个单位,X_2 减少5个单位,那么 Y 的期望变化又是多少?

5.1 在表中添加“ * ”(5%)和“ * * ”(1%)以表示系数的统计显著性程度。

5.2 计算每个回归的 $\bar{R}^2$。

5.3 利用第(1)列中的回归结果:

*a. 平均来讲,有大学学历的工人是否比只有高中文凭的工人挣得多? 多多少? 根据这个回归所估计的收入差异在5%的水平下在统计上是否是显著的?

b. 平均来讲,男性是否比女性挣得多? 多多少? 根据这个回归所估计的收入差异在5%的水平下在统计上是否是显著的?

5.4 利用第(2)列的回归结果:

a. 年龄是否是收入的一个重要的决定性因素? 请解释说明。

b. Sally 是一位29岁的女大学毕业生。Betty 是一位34岁的女大学毕业生。预测 Sally 和 Betty 的收入,并构造她们收入之间期望差异的95%的置信区间。

5.5 利用第(3)列中的回归结果:

a. 收入是否存在重要的地域差异?

b. 为什么回归中遗漏了回归因子 *West*? 如果它被包含进来,会发生什么情况?

*c. Juanita 是一位来自于南部的28岁的女大学毕业生,Molly 是一位来自于西部的28岁的女大学毕业生,Jennifer 是一位来自于中西部的28岁的女大学毕业生。

ci. 构造 Juanita 和 Molly 之间期望收入之差的95%的置信区间。

cii. 计算 Juanita 和 Jennifer 的期望收入之差。

ciii. 解释说明如何构造 Juanita 和 Jennifer 之间期望收入之差的95%的置信区间。(提示:如果你在回归中包含了 *West* 并剔除了 *Midwest*,会发生什么情况?)

5.6 重新估计第(2)列中所显示的回归,这次使用1992年的数据(4 000个观察值是从1993年3月的CPS中随机选择的,利用消费者价格指数将其转换为1998年的美元),结果是:

$$\widehat{AHE} = \underset{(0.98)}{0.77} - \underset{(0.18)}{5.29}\,College - \underset{(0.18)}{2.59}\,Female + \underset{(0.03)}{0.40}\,Age, SER = 5.85, \bar{R}^2 = 0.21$$

将这个回归与第(2)列所显示的对1998年数据的回归相比较,*College* 的系数在统计上是否存在显著的变化?

*5.7 评价下面的陈述:“在所有的回归中,*Female* 的系数是负的、大的,而且在统计上是显著的。这提供了在美国劳动力市场上存在性别歧视的强统计证据。”

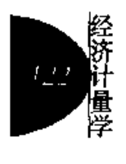
5.8 考虑回归模型:$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + u_i$。用5.8节中的“方法2”变换这个回归,这样你可以用 t 统计量进行检验:

a. $\beta_1 = \beta_2$;

b. $\beta_1 + a\beta_2 = 0$,这里 a 是个常数;

c. $\beta_1 + \beta_2 = 1$。(提示:你必须重新定义回归中的因变量)

5.9 附录5.3证明了经验规则 F 统计量的两个公式:公式(5.38)和公式(5.39)。证明:这两个公式是等价的。



附录5.1 表达式(5.1)的推导

本附录介绍了表达式(5.1)中的遗漏变量偏差公式的推导。附录4.3中的公式(4.51)说明了:

5.7 节中的 F 统计量。

本附录描述了检验联合假设的其他两种方法，即当你只有回归分析结果表时，可以使用这两种方法。第一种方法是 Bonferroni 检验，它是基于 Bonferroni 不等式的一个非常一般的检验方法的应用。第二种方法是经验规则 F 统计量，它是多元回归的一个很专门的方法，只有在误差项是同方差的时，这种方法在理论上才是合理的。经验规则 F 统计量是用仅适用于同方差的标准误计算的 t 统计量所对应 F 统计量。

Bonferroni 检验

Bonferroni 检验是基于 t 统计量对单个假设的联合假设的检验，即 Bonferroni 检验是 5.7 节中介绍的效果较好的"每次一个"方法的 t 统计量检验。基于临界值 $c>0$，联合零假设 $\beta_1=\beta_{1,0}$ 和 $\beta_2=\beta_{2,0}$ 的 Bonferroni 检验(Bonferroni test)的规则是：

$$\text{如果}|t_1|\leqslant c\text{ 且}|t_2|\leqslant c\text{，那么就接受；否则就拒绝} \tag{5.35}$$

(Bonferroni"每次一个"方法的 t 统计量检验)

这里，t_1 和 t_2 分别是检验 β_1 和 β_2 的约束条件的 t 统计量。

选择临界值 c 可以采用这样的技巧：当零假设为真时，"每次一个"的检验拒绝零假设的概率不超过所想要的显著性水平，比如说 5%。选择临界值 c 可以通过使用 Bonferroni 的不等式完成，这种方法既考虑到了同时检验两个约束条件，同时又考虑到了 t_1 和 t_2 之间任何可能的相关性。

Bonferroni 的不等式

Bonferroni 不等式是概率论的一个基本结果。设 A 和 B 为两个事件。设 $A\cap B$ 为"A 和 B 同时发生"的事件(A 和 B 的交集)，记 $A\cup B$ 为"A 和 B 至少有一个发生"的事件(A 和 B 的并集)，那么，$\Pr(A\cup B)=\Pr(A)+\Pr(B)-\Pr(A\cap B)$。由于 $\Pr(A\cap B)\geqslant 0$，由此可得 $\Pr(A\cup B)\leqslant\Pr(A)+\Pr(B)$。这个不等式也可转换为 $1-\Pr(A\cup B)\geqslant 1-[\Pr(A)+\Pr(B)]$。设 A^c 和 B^c 分别是 A 和 B 的补集，即"非 A"和"非 B"事件。由于 $A\cup B$ 的补集为 $A^c\cap B^c$，因此有 $1-\Pr(A\cup B)=\Pr(A^c\cap B^c)$，这就得到了 Bonferroni 不等式 $\Pr(A^c\cap B^c)\geqslant 1-[\Pr(A)+\Pr(B)]$。

现在设 A 为事件 $|t_1|>c$，设 B 为事件 $|t_2|>c$，那么由不等式 $\Pr(A\cup B)\leqslant\Pr(A)+\Pr(B)$ 可以得到：

$$\Pr(|t_1|>c\text{ 和}|t_2|>c\text{ 至少有一个发生})\leqslant\Pr(|t_1|>c)+\Pr(|t_2|>c) \tag{5.36}$$

Bonferroni 检验

由于事件"$|t_1|>c$ 和 $|t_2|>c$ 至少有一个发生"是"每次一个"检验的拒绝域，因此，表达式(5.36)提供了选择临界值 c 的一种方法，这种方法使得"每次一个"的 t 统计量在大样本条件下实现想要的显著性水平。在大样本的零假设下，$\Pr(|t_1|>c)=\Pr(|t_2|>c)=\Pr(|Z|>c)$，因此，表达式(5.36)意味着在大样本条件下"每次一个"的检验在零假设下拒绝零假设的概率是：

$$\Pr_{H_0}(\text{"每次一个"的检验拒绝})\leqslant 2Pr(|Z|>c) \tag{5.37}$$

表达式(5.37)中的不等式提供了选择临界值 c 的一种方法，该方法在零假设下拒绝的概率等于想要达到的显著性水平。Bonferroni 方法可以扩展到两个以上系数的情形。如果在零假设下有 q 个约束条件，那么用 q 来代替表达式(5.37)中右边的因子 2。



表 5—3 给出了在不同的显著性水平下，$q=2$、$q=3$ 和 $q=4$ 时，"每次一个"的 Bonferroni 检验的临界值 c。例如，假设想要的显著性水平为 5%，且 $q=2$，根据表 5—3，临界值 c 为

2.241。这个临界值是标准正态分布的1.25%百分位数,所以$Pr(|Z|>2.241)=2.5\%$。因此,表达式(5.37)告诉我们,在大样本中,表达式(5.35)中"每次一个"的检验在零假设下至多有5%的机会拒绝零假设。

表5—3　　检验联合假设的"每次一个"的t统计量的Bonferroni临界值c

约束条件个数(q)	显著性水平		
	10%	5%	1%
2	1.960	2.241	2.807
3	2.128	2.394	2.935
4	2.241	2.498	3.023

表5—3中的临界值比检验单个约束条件的临界值大。例如,在$q=2$时,如果至少有一个t统计量的绝对值大于2.241,那么"每次一个"的检验将拒绝零假设。这个临界值大于1.96,因为如同在5.7节中所讨论的,通过考察两个t统计量,你得到了第二次机会拒绝联合零假设,所以上述的临界值被做了适当的修正。

如果单个的t统计量是以异方差稳健的标准误为基础的,那么不论是否存在异方差,Bonferroni检验都是有效的,但是如果t统计量是基于仅适用于同方差的标准误的,那么只有在同方差条件下Bonferroni检验才是有效的。

在考试分数例子中的应用

检验公式(5.18)中关于考试成绩和平均每个学生费用的真实系数的联合零假设的t统计量分别是$t_1=-0.60$和$t_2=2.43$。尽管$|t_1|<2.241$,但因为$|t_2|>2.241$,所以用Bonferroni检验,我们能够在5%的显著性水平下拒绝联合零假设。然而,t_1和t_2的绝对值都小于2.807,因此用Bonferroni检验,我们不能在1%的显著性水平拒绝联合零假设。相反,使用5.7节中的F统计量,我们却能够在1%的显著性水平下拒绝这个假设。

经验规则F统计量

经验规则F统计量是在两个回归的残差平方和的基础上用一个简单的公式计算得到的。在第一个回归中,也即被称为有约束的回归(restricted regression)中,零假设是被强制为真的。当零假设是表达式(5.20)中的类型时,其中所有的假设值都是零,有约束的回归是将这些系数设定为0的回归,也就是说,相关的回归因子被排除在回归之外。在第二个回归中,也即被称为无约束的回归(unrestricted regression)中,备择假设是允许为真的。如果无约束的回归中的残差平方和比有约束的回归中的残差平方和足够地小,那么检验将拒绝零假设。

经验规则F统计量(rule-of-thumb F-statistic)由下式给出:

$$F=\frac{(SSR_{restricted}-SSR_{unrestricted})/q}{SSR_{unrestricted}/(n-k_{unrestricted}-1)} \tag{5.38}$$

这里,$SSR_{restricted}$是有约束的回归的残差平方和,$SSR_{unrestricted}$是无约束的回归的残差平方和,q是零假设下的约束条件个数,$k_{unrestricted}$是无约束的回归中回归因子的个数。经验规则F统计量的另一个等价的公式,是以两个回归的R^2为基础计算的,即:

$$F=\frac{(R^2_{unrestricted}-R^2_{restricted})/q}{(1-R^2_{unrestricted})/(n-k_{unrestricted}-1)} \tag{5.39}$$

如果误差项是同方差的,那么用公式(5.38)计算的经验规则F统计量和5.7节中所用



到的 F 统计量之间的差异会随着样本容量 n 的增加而逐渐消失。因而，如果误差项是同方差的，那么在大样本中，经验规则 F 统计量在零假设下的抽样分布是 $F_{q,\infty}$。

这些经验规则公式是易于计算的，而且根据有约束的和无约束的回归对数据的拟合程度如何有直观的解释。遗憾的是，只有当误差项是同方差的时，它们才是有效的。因为同方差是一种极特殊的情况，我们不能指望在经济数据应用中或在更一般的社会科学数据应用中出现同方差的情况，因此经验规则 F 统计量不是 5.7 节中异方差稳健的 F 统计量的一个满意的替代。

在考试成绩和学生—教师比案例中的应用

在控制 *PctEL* 的情况下，为了检验关于 *STR* 和 *Expn* 的总体系数均为零的零假设，我们需要计算有约束的和无约束的回归中的 *SSR*（或 R^2）。无约束的回归中含有回归因子 *STR*、*Expn* 和 *PctEL*，由公式(5.18)给出；它的 R^2 是 0.4366，即 $R^2_{unrestricted}=0.4366$。对有约束的回归施加了关于 *STR* 和 *Expn* 的真实系数均为零的联合零假设，即在零假设下，*STR* 和 *Expn* 不进入总体回归，尽管 *PctEL* 进入总体回归（零假设没有约束 *PctEL* 的系数）。用 OLS 所估计的有约束的回归是：

$$\widehat{TestScore} = \underset{(1.0)}{664.7} - \underset{(0.032)}{0.671} \times PctEL, R^2=0.4149 \tag{5.40}$$

因此 $R^2_{restricted}=0.4149$。约束条件个数为 $q=2$，观察值个数为 $n=420$，无约束的回归中回归因子的个数为 $k=3$。用公式(5.39)计算的经验规则 F 统计量是：

$$F=[(0.4366-0.4149)/2]/[(1-0.4366)/(420-3-1)]=8.01$$

由于 8.01 大于 1% 的临界值 4.61，因此利用这个经验规则方法，零假设在 1% 的显著性水平下被拒绝。

这个例子阐明了经验规则 F 统计量的优缺点。它的优点是它可以用计算器进行计算。它的缺点是经验规则 F 统计量的值可能完全不同于 5.7 节中所用的异方差稳健的 F 统计量：检验这个联合假设的异方差稳健的 F 统计量的值为 5.43，与不太可靠的仅适用于同方差的经验规则值 8.01 之间有较大的不同。

第6章

非线性回归函数

在第4章和第5章中，总体回归函数被假定是线性的，换句话说，总体回归函数的斜率是常数，因此，X 的单位变化对 Y 的影响并不依赖于 X 本身的取值。但是，如果 X 的变化对 Y 的影响依赖于一个或多个自变量的取值，那么总体回归函数是什么样的呢？如果是这样的话，总体回归函数就是非线性的。

本章提出了两组发现非线性总体回归函数并对其进行建模的方法。当一个自变量 X_1 的变化对 Y 的影响依赖于 X_1 本身的取值时，第一组方法是有用的。比如，在班级规模已经比较小的情况下，将班级规模指标即每个教师分担的学生数减少一人，比在那些班级规模已经很大，以至于教师除了控制班级秩序以外几乎不能做什么的情况下，会产生更大的影响。如果是这样的话，考试成绩(Y)将是学生—教师比(X_1)的非线性函数，这里当 X_1 很小时，这个函数是比较陡峭的。图6—1显示了具有这个特征的非线性函数的一个例子。图6—1(a)中的线性总体回归函数具有一个常数斜率，而图6—1(b)中的非线性总体回归函数当 X_1 值很小时的斜率比当 X_1 值很大时的斜率陡峭。这个第一组的方法在6.2节中论述。

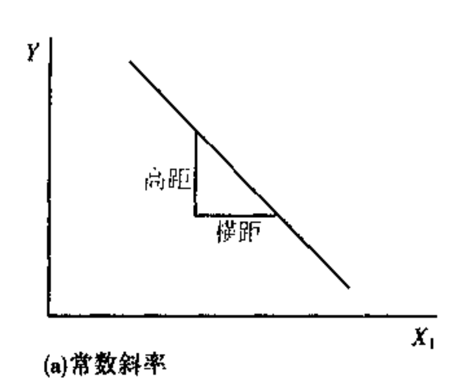
当 X_1 的变化对 Y 的影响依赖于另一个自变量时，比如说 X_2，第二组方法是有用的。例如，仍在学习英语的学生尤其可能会受益于有较多的一对一的关注。如果是这样的话，降低学生—教师比对考试成绩的影响，在具有许多仍在学习英语的学生的地区，比在具有较少的英语学习者的地区的影响大。在这个例子中，学生—教师比(X_1)的下降对考试成绩(Y)的影响依赖于该地区中的英语学习者的百分比(X_2)。如图6—1(c)所示，这类总体回归函数的斜率依赖于 X_2 的值。这个第二组方法在6.3节中论述。

在本章的模型中，总体回归函数是自变量的非线性函数，也就是说，条件期望 $E(Y_i|X_{1i},\cdots,X_{ki})$ 是一个或多个 X 变量的非线性函数。尽管它们是变量 X 的非线性函数，但这些模型却是总体回归模型未知系数(或参数)的线性函数，进而是第5章中多元回归模型的形式。因此，非线性回归函数的未知系数能够用普通最小二乘法(OLS)以及第5章中的方法进行估计和检验。

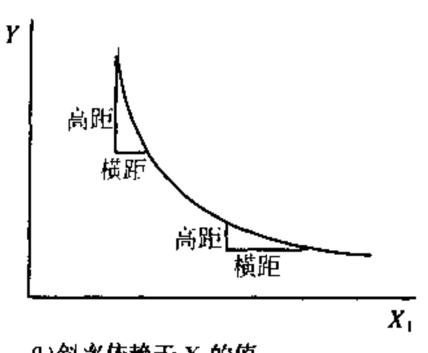


6.1节和6.2节介绍了具有单个自变量回归情形下的非线性回归函数，而6.3节将其扩展到两个自变量的情形。为了简单起见，6.1节到6.3节中的实证例子省略了额外控制

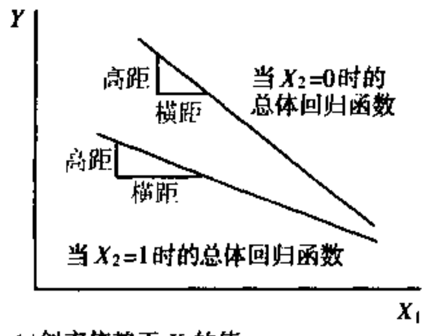




(a)常数斜率



(b)斜率依赖于 X_1 的值



(c)斜率依赖于 X_2 的值

注:在图 6—1(a)中,总体回归函数具有常数斜率。在图 6—1(b)中,总体回归函数的斜率依赖于 X_1 的值。在图 6—1(c)中,总体回归函数的斜率依赖于 X_2 的值。

图 6—1 具有不同斜率的总体回归函数

变量。然而在实际中,通过引入控制变量,进而在有控制遗漏变量偏差的模型中分析非线性回归函数是非常重要的。在 6.4 节,在学生特征保持不变的情况下,当我们进一步观察考试成绩与学生—教师比之间可能的非线性关系时,我们将非线性回归函数和额外的控制变量结合在一起。

6.1 非线性回归函数建模的一般策略

这节展开介绍了非线性总体回归函数建模的一般策略。在这种策略中,非线性模型是作为多元回归模型的扩展形式,因此可用第 5 章中的工具进行估计和检验。我们首先回到加利福尼亚州考试成绩数据的例子中,并考虑考试成绩与地区收入之间的关系。

6.1.1 考试成绩与地区收入

在第 5 章,我们发现学生的经济背景是解释标准化考试成绩的一个重要因素。在上一章的分析中,我们使用了两个经济背景变量(享受午餐补助计划的学生百分比和享受收入援助计划的地区家庭的百分比)来测度地区中来自贫困家庭的学生的比率。另外一个不同的测度经济背景的更广泛应用的指标是学区中的平均年人均收入("地区收入")。加利福尼亚州的数据集中包含有 1998 年当年以千美元测度的地区收入数据。该样本数据中包含了一个较宽范围的收入水平数据:对我们样本中的 420 个学区而言,学区收入中位数是 13.7(即人均收入 13 700美元),它从 5.3(人均 5 300 美元)变动到 55.3(人均 55 300 美元)。

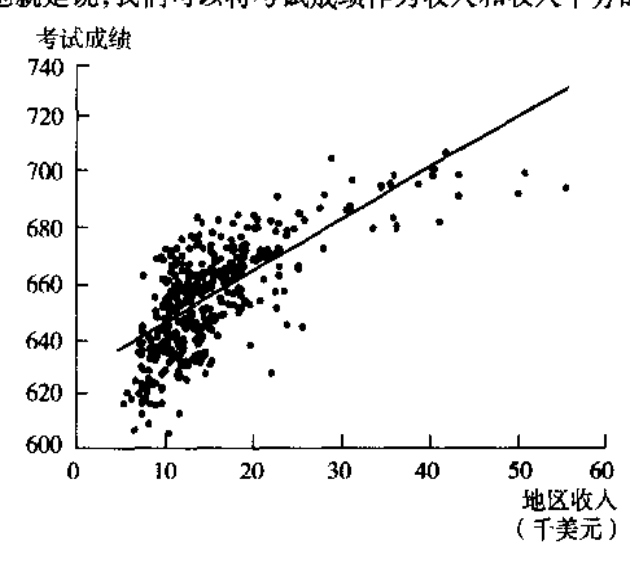
图 6—2 反映了加利福尼亚州数据集中一个五年级学生的考试成绩与地区收入的散点

图,同时附加了联系这两个变量的普通最小二乘回归线。考试成绩和平均收入是强正相关的、相关系数为0.71。来自富裕地区的学生在考试上的表现要比来自贫困地区的学生好。但这个散点图有一个特殊性:当收入很低(10 000 美元以下)或很高(40 000 美元以上)时,大多数点在普通最小二乘回归线的下方,而当收入介于15 000 美元~30 000 美元之间时则在回归线的上方。在考试成绩与收入之间的关系中似乎还有一些没有被线性回归捕捉到的凸性关系。

简而言之,地区收入与考试成绩之间的关系看上去似乎并不是一条直线,或者确切地说,它是非线性的曲线。非线性函数是个斜率不是常数的函数:如果函数$f(X)$的斜率对于所有的X值都是相同的,那么$f(X)$是线性的,但如果斜率依赖于X的值,那么$f(X)$是非线性的。



如果直线不足以描述地区收入与考试成绩之间的关系,那么用什么来描述它们呢?设想画一条拟合图6—2中的点的曲线,这条曲线对于低的地区收入将会是陡峭的,而后随着地区收入变高曲线将会变得平坦。数学上,近似这种曲线的一种方法就是用二次函数模型来模拟这个关系,也就是说,我们可以将考试成绩作为收入和收入平方的函数来建模。



注:在考试成绩和地区收入之间存在正相关关系(相关系数=0.71),但线性OLS回归线不能充分地描述这两个变量之间的关系。

图6—2 具有线性OLS回归函数的考试成绩与地区收入的散点图

考试成绩与收入之间关系的二次总体回归模型的数学表达式为:

$$TestScore_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2 + u_i \tag{6.1}$$

其中,β_0,β_1和β_2是系数,$Income_i$是第i个地区的收入,$Income_i^2$是第i个地区的收入平方,而u_i是误差项,按照惯例,它代表决定考试成绩的所有其他因素。公式(6.1)被称为二次回归模型(quadratic regression model),因为总体回归函数$E(TestScore_i \mid Income_i) = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$是自变量$Income$的二次函数。

如果已知公式(6.1)中的总体系数β_0,β_1和β_2,那么就可以在一个地区的平均收入基础上预测该地区的考试成绩。但是这些总体系数都是未知的,因此必须用样本数据进行估计。

求出能够最佳拟合图6—2中数据的二次函数的系数,乍看起来可能很难。然而,如果将公式(6.1)与重要概念5.2中的多元回归模型进行比较,那么就会发现,公式(6.1)实际上是一种具有两个回归因子的多元回归模型:第一个回归因子是$Income$,第二个回归因子是$Income^2$。因此,在将回归因子定义为$Income$和$Income^2$之后,公式(6.1)中的非线性模型只



变化计算起来更复杂，因为它依赖于自变量的值。

非线性总体回归函数的一般表达式[1]。本章中所考虑的非线性总体回归模型如下式：

$$Y_i = f(X_{1i}, X_{2i}, \cdots, X_{ki}) + u_i, i = 1, \cdots, n, \tag{6.3}$$

其中 $f(X_{1i}, X_{2i}, \cdots, X_{ki})$ 是总体**非线性回归函数**(nonlinear regression function)，它是自变量 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 的一个可能的非线性函数，u_i 是误差项。例如，在公式(6.1)的二次回归模型中，只有一个自变量，因此 X_1 是 *Income*，总体的回归函数为 $f(Income_i = \beta_0 + \beta_1 Income_i + \beta_2 Income_i^2$。

因为总体的回归函数是给定 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 下 Y_i 的条件期望，所以在公式(6.3)中，我们考虑了这个条件期望是 $X_{1i}, X_{2i}, \cdots, X_{ki}$ 一个非线性函数的可能性，即 $E(Y_i | X_{1i}, X_{2i}, \cdots, X_{ki}) = f(X_{1i}, X_{2i}, \cdots, X_{ki})$，其中 f 可能是个非线性函数。如果总体回归函数是线性的，那么 $f(X_{1i}, X_{2i}, \cdots, X_{ki}) = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$，公式(6.3)变为重要概念5.2中的线性回归模型。不过，公式(6.3)还考虑了非线性回归函数。

X_1 的变化对 Y 的效应。如5.2节中所讨论的，在 $X_2, \cdots, X_k$ 保持不变的情况下，X_1 变化 ΔX_1 对 Y 的效应就是当自变量取 $X_1 + \Delta X_1, X_2, \cdots, X_k$ 时 Y 的期望值和当自变量取 $X_1, X_2, \cdots, X_k$ 时 Y 的期望值之差。这两个期望值之差，如 ΔY，是在其他变量 $X_2, \cdots, X_k$ 保持不变的情况下，当 X_1 改变一定量 ΔX_1 时，平均来看在总体中 Y 所发生的变化。在公式(6.3)的非线性回归模型中，这对 Y 的效应为 $\Delta Y = f(X_1 + \Delta X_1, X_2, \cdots, X_k) - f(X_1, X_2, \cdots, X_k)$。

由于回归函数 f 是未知的，因此 X_1 的变化对 Y 的总体效应也是未知的。为了估计总体效应，首先要估计总体回归函数。一般地说，将这个估计函数表示为 $\hat{f}$，这种估计函数的一个例子是公式(6.2)中所估计的二次回归函数。X_1 的变化对 Y 的估计效应(记为 $\Delta\hat{Y}$)是当自变量取 $X_1 + \Delta X_1, X_2, \cdots, X_k$ 值时 Y 的预测值和当自变量取 $X_1, X_2, \cdots, X_k$ 值时 Y 的预测值之差。

计算 X_1 的变化对 Y 的期望效应的方法在重要概念6.1中总结。

在考试成绩和收入例子中的应用。根据公式(6.2)中所估计的二次回归函数，与地区收入的1 000美元的变化相联系的考试成绩的期望变化是多少？由于那个回归函数是二次的，因此这个效应依赖于初始的地区收入。所以我们考虑两种情况：地区收入从10增加到11(即人均收入从10 000美元增加到11 000美元)和地区收入从40增加到41。

为了计算与收入从10变化到11有关的 $\Delta\hat{Y}$，我们可将公式(6.6)中的一般表达式应用于二次回归模型，这样就得到：

$$\Delta\hat{Y}(\hat{\beta}_0 + \hat{\beta}_1 \times 11 + \hat{\beta}_2 \times 11^2) - (\hat{\beta}_0 + \hat{\beta}_1 \times 10 + \hat{\beta}_2 \times 10^2) \tag{6.4}$$

其中，$\hat{\beta}_0$，$\hat{\beta}_1$ 和 $\hat{\beta}_2$ 是OLS估计量。

公式(6.4)第一个括号中的项是当 *Income* = 11时 Y 的预测值，而在第二个括号中的项是当 *Income* = 10时 Y 的预测值。这些预测值是用公式(6.2)中系数的OLS估计值进行计算的。因此，当 *Income* = 10时，考试成绩的预测值为 $607.3 + 3.85 \times 10 - 0.0423 \times 10^2 = 641.57$；当 *Income* = 11时，预测值为 $607.3 + 3.85 \times 11 - 0.0423 \times 11^2 = 644.53$。这两个预测值之差为 $\Delta\hat{Y} = 644.53 - 641.57 = 2.96$ 分，也就是说，平均收入为11 000美元的地区与平均收入为10 000美元的地区之间的考试成绩的预测差为2.96分。

[1] 术语"非线性回归"适用于两个概念上不同的模型族。在第一个模型族中，总体回归函数是变量 X 的非线性函数，但却是未知参数(β 系数)的线性函数；在第二个模型族中，总体回归函数是未知参数的非线性函数，且可能是或可能不是 X 变量的非线性函数。本章中的模型都属于第一个模型族。在第9章中，当我们开始讨论二元因变量的回归时，我们会遇到来自第二个模型族的模型。



在第二种情况下，当收入从40 000美元变化到41 000美元时，公式(6.4)中的预测值之差为$\Delta\hat{Y}=(607.3+3.85\times41-0.0423\times41^2)-(607.3+3.85\times40-0.0423\times40^2)=694.04-693.62=0.42$分。这样，在初始收入为10 000美元的情况下，与1 000美元的收入变化相联系的预测考试成绩的变化比初始收入为40 000美元的情况下的变化大(预测变化为2.96分与0.42分)。换句话说，图6—3中所估计的二次回归函数的斜率在低收入值处(如10 000美元)比在较高的收入值处(如40 000美元)陡峭。

重要概念6.1

在非线性回归模型(6.3)中X_1的变化对Y的期望效应

在$X_2,\cdots,X_k$保持不变的情况下，与X_1变化ΔX_1相联系的Y的期望变化ΔY，是在$X_2,\cdots,X_k$保持不变的情况下，X_1变化前后总体回归函数值之差，也就是说，Y的期望变化是：

$$\Delta Y=f(X_1+\Delta X_1,X_2,\cdots,X_k)-f(X_1,X_2,\cdots,X_k) \tag{6.5}$$

这个未知总体差值的估计量，是这两种情况下的预测值之差。设$\hat{f}(X_1,X_2,\cdots,X_k)$为基于总体回归函数的估计量$\hat{f}$的$Y$的预测值，那么$Y$的预测变化为：

$$\Delta\hat{Y}=\hat{f}(X_1+\Delta X_1,X_2,\cdots,X_k)-\hat{f}(X_1,X_2,\cdots,X_k) \tag{6.6}$$

估计效应的标准误。X_1的变化对Y的效应的估计量，依赖于总体回归函数的估计量$\hat{f}$，这个估计量随样本的变化而不同。因此，估计效应中包含抽样误差。一种量化与估计效应有关的抽样不确定性的方法，是计算真实总体效应的置信区间。为了计算真实总体效应的置信区间，我们需要计算公式(6.6)中$\Delta\hat{Y}$的标准误。

当回归函数是线性的时，很容易计算$\Delta\hat{Y}$的标准误。X_1的变化的估计效应为$\hat{\beta}_1\Delta X_1$，因此所估计的变化的95%的置信区间是$\hat{\beta}_1\Delta X_1\pm1.96SE(\hat{\beta}_1)\Delta X_1$。

在本章的非线性回归模型中，$\Delta\hat{Y}$的标准误可用5.8节中所介绍的涉及多个系数的检验单个约束条件的工具进行计算。为了具体说明这种方法，考虑在公式(6.4)中与收入从10到11的变化相联系的考试成绩的估计变化，它是$\Delta\hat{Y}=\hat{\beta}_1\times(11-10)+\hat{\beta}_2\times(11^2-10^2)=\hat{\beta}_1+21\hat{\beta}_2$。因而，预测变化的标准误为：

$$ST(\Delta\hat{Y})=SE(\hat{\beta}_1+21\hat{\beta}_2) \tag{6.7}$$

因此，如果我们能够计算$\hat{\beta}_1+21\hat{\beta}_2$的标准误，那么我们就已经计算出了$\Delta\hat{Y}$的标准误。使用标准的回归软件可有两种方法来计算标准误，它们对应于5.8节中检验单个约束条件对多个系数的两种方法[①]。

第一种方法就是用5.8节中的“方法1”，这种方法就是要计算检验假设$\hat{\beta}_1+21\hat{\beta}_2=0$的$F$统计量。那么$\Delta\hat{Y}$的标准误由下式给出[②]：

$$SE(\Delta\hat{Y})=\frac{|\Delta\hat{Y}|}{\sqrt{F}} \tag{6.8}$$

当公式(6.8)应用于公式(6.2)中的二次回归时，检验假设$\hat{\beta}_1+21\hat{\beta}_2=0$的$F$统计量是$F=$

① 这两种方法是用不同的方式使用回归软件来执行16.2节中所提出的预测效应标准误的一般表达式。

② 公式(6.8)是根据F统计量是检验这个假设的t统计量的平方推导出来的，即：$F=t^2=[(\hat{\beta}_1+21\hat{\beta}_2)/SE(\hat{\beta}_1+21\hat{\beta}_2)]^2=[\Delta\hat{Y}/SE(\Delta\hat{Y})]^2$，进而解出$SE(\Delta\hat{Y})$。

299.94。由于 $\Delta \hat{Y}=2.96$，应用公式(6.8)，得出 $SE(\Delta \hat{Y})=2.96/\sqrt{299.94}=0.17$，因而，$Y$ 的期望值的变化的95%的置信区间为 $2.96\pm1.96\times0.17$，即为(2.63,3.29)。

第二种方法就是用5.8节中的"方法2"，这种方法需要变换回归因子，使得在所变换的回归中有一个系数是 $\beta_1+21\beta_2$。这个变换留作练习(见练习6.4)。

在非线性设定下对系数解释的评论。在第5章的多元回归模型中，回归系数拥有自然的解释。例如，β_1 是在保持其他回归因子不变的情况下，与 X_1 的变化有关的 Y 的期望变化。但是，正如我们已看到的，非线性模型一般并非如此。也就是说，将公式(6.1)中的 β_1 看做是在保持地区收入的平方不变的情况下改变地区收入所产生的影响，这种认识是没有多大帮助的。这意味着，在非线性模型中，回归函数最好是通过画图和计算改变一个或多个自变量对 Y 的预测效应进行解释。



6.1.3 用多元回归建立非线性模型的一般方法

本章中所采用的建立非线性回归函数模型的一般方法包括五个要素：

1. 识别可能的非线性关系。最好的办法就是利用经济理论和你所了解的关于应用方面的知识，建议一种可能的非线性关系。甚至在分析数据之前，先想想联系 Y 和 X 的回归函数的斜率是否很合理地依赖于 X 的值或其他自变量的值。为什么会存在这样的非线性依赖性？建议一种什么样的非线性形式？例如，考虑一下11岁小学生的班级，将班级规模由18人削减到17人可能会比由30人削减到29人有更大的影响，看看这个例子能给我们什么启示。

2. 设定一个非线性函数并用OLS估计其参数。6.2节和6.3节含有可用OLS方法估计的各种不同的非线性回归函数。学懂这两节的内容后，你就会理解这些函数中每个函数的特征。

3. 确认所确定的非线性模型是否比线性模型有所改进。并不一定因为你想象的函数是非线性的，它确实就是非线性的！你必须在经验的基础上证明你所建立的非线性模型是否是适合的。大多数情况下，你可以用 t 统计量和 F 统计量来检验总体回归函数是线性的零假设，对应于它是非线性的备择假设。

4. 绘制所估计的非线性回归函数。所估计的回归函数是否很好地描述了数据的分布？看一看图6—2和图6—3，这两个图表明了二次模型对数据的拟合要比线性模型好。

5. 估计 X_1 的变化对 Y 的效应。最后一步是利用重要概念6.1中的方法，用所估计的回归方程计算一个或多个回归因子 X 的变化对 Y 的效应。

6.2 单个自变量的非线性函数

本节提出了建立非线性回归函数模型的两种方法。为了简单起见，我们逐步展开这些只涉及一个自变量 X 的非线性回归函数的方法。不过，在6.4节中我们会看到，这些模型经过修正，就可以包含多个自变量。

本节所讨论的第一种方法是多项式回归，这种回归是上一节中用来建立考试成绩和收入之间关系的二次回归模型的推广。第二种方法使用 X 和/或 Y 的对数。尽管这些方法是分别被提出来的，但它们可以组合起来使用。



6.2.1 多项式

设定非线性回归函数的一种方法就是使用 X 的多项式。一般地说，设 r 表示回归中包

含的 X 的最高次幂。r 次**多项式回归模型**(polynomial regression model)为:

$$Y_i=\beta_0+\beta_1X_i+\beta_2X_i^2+\cdots+\beta_rX_i^r+u_i \tag{6.9}$$

当 $r=2$ 时,公式(6.9)就是6.1节中所讨论的二次回归模型。当 $r=3$ 时,所包含的 X 的最高次幂是 X^3,公式(6.9)被称为**三次回归模型**(cubic regression model)。

多项式回归模型与第5章中的多元回归模型很相似,其差别在于第5章中的回归因子是不同的自变量,而这里的回归因子是同一个自变量 X 的不同次幂,即回归因子为 X,X^2,X^3 等,因此,多元回归中所提出的估计和推断的技术都可以应用在这里,尤其是公式(6.9)中的未知系数 $\beta_0,\beta_1,\cdots,\beta_r$,可以用 Y_i 对 $X_i,X_i^2,\cdots,X_i^r$ 回归的 OLS 方法进行估计。



检验总体回归函数是线性的零假设。如果总体回归函数是线性的,那么二次项和较高次的项就不会进入到总体回归函数中。因此,回归是线性的零假设(H_0)以及它是 r 次多项式的备择假设(H_1)对应于:

$$H_0:\beta_2=0,\beta_3=0,\cdots,\beta_r=0;H_1:至少有一个\beta_j\neq0,j=2,\cdots,r \tag{6.10}$$

总体回归函数是线性的零假设,对应于它是 r 次多项式的备择假设,可以通过表达式(6.10)中 H_0 对 H_1 的检验来完成。由于 H_0 是个联合零假设,其中对总体多项式回归模型的系数有 $q=r-1$ 个约束条件,它可以用5.7节中所描述的 F 统计量进行检验。

应该使用几次多项式?也就是说,多项式回归中应该包含 X 的多少次幂?答案要在灵活性和统计精确度之间进行权衡。增加次数 r 会使回归函数拥有更大的灵活性,并使之与数据模式更好地相匹配。一个 r 次多项式在其图形中的弯曲部分(即拐点)可以达到 $r-1$ 个,但是增加 r 意味着增加了更多的回归因子,这会降低估计系数的精确度。

因此,在回归中包含多少项的问题的答案,就是在非线性回归建模中应该包含足够多的项,但也不能过多。不幸的是,这个答案在实践中并没有多大用处。

一种较为实际的确定多项式次数的方法,就是判断公式(6.9)中与 r 的最大值相对应的系数是否为零。如果是的话,那么就可以从回归中剔除这些项。这个方法被称为序贯假设检验,因为每个假设被依次有序地进行检验。序贯假设检验的步骤总结如下:

1. 选取 r 的最大值,并估计出带有该 r 次项的多项式回归。

2. 用 t 统计量检验 X^r 的系数(公式(6.9)中的 β_r)为零的假设。如果这个假设被拒绝,那么 X^r 应属于这个回归,因此使用 r 次多项式。

3. 如果在第2步中你没有拒绝 $\beta_r=0$,那么就将 X^r 从回归中剔除,并继续估计 $r-1$ 次多项式回归。检验 X^{r-1} 的系数是否为零。如果拒绝,那么就使用 $r-1$ 次多项式。

4. 如果在第3步中你没有拒绝 $\beta_{r-1}=0$,那么就继续这个步骤,直到多项式中最高次幂的系数在统计上是显著的。

这种方法有个缺失的成分:多项式的初始次数 r。在很多涉及经济数据的应用中,非线性函数是平滑的,即它们没有突然的跳跃或"尖峰值"。如果是这样的话,那么选择一个小一点的多项式的最大阶数是适合的,比如说2,3或4,即第1步从 $r=2$、$r=3$ 或 $r=4$ 开始①。

在考试成绩和地区收入案例中的应用。关于地区收入和考试成绩之间关系的三次回归函数是:

$$\widehat{TestScore}=\underset{(5.1)}{600.1}+\underset{(0.71)}{5.02}\,Income-\underset{(0.029)}{0.096}\,Income^2+\underset{(0.00035)}{0.00069}\,Income^3,\ \bar{R}^2=0.555 \tag{6.11}$$

$Income^3$ 的 t 统计量为1.97,因此,与回归函数是三次的这一备择假设相对应的回归函数是



① 选择 r 的一种不同的方法就是使用"信息准则",它在第12章"时间序列分析"部分中介绍。实际上,信息准则方法和这里介绍的序贯假设检验方法经常会得出相似的结果。

二次的这一零假设,在5%的显著性水平下被拒绝。此外,检验 $Income^2$ 和 $Income^3$ 的系数均为零的联合零假设的 F 统计量为37.7,p 值小于0.01%,因此对应于回归函数是三次的备择假设,它是线性的零假设被拒绝。

多项式回归模型系数的解释。对于多项式回归的系数,没有很简单的解释。解释多项式回归的最好方法就是画出所估计的回归函数图形,并计算出一个或多个 X 值的变化对 Y 的估计效应。

6.2.2 对 数



设定非线性回归函数的另一种方法就是使用 X 和/或 Y 的自然对数。对数将变量的绝对变化转换成百分比变化,而且许多关系是用百分比变化来表示的。下面是一些例子:

■ 3.5 节研究了男女大学毕业生之间的工资差异。在那个讨论中,工资差异是用美元来测度的。可是,当它们用百分比的形式表示时,会更易于比较不同职业和不同时间的工资差异。

■ 在6.1 节,我们发现地区收入和考试成绩之间是非线性关系。若用百分比变化表示,这个关系会是线性的吗?也就是说,对不同收入水平而言,地区收入的1%的变化——而不是1 000美元——会与考试成绩的变化密切联系吗(在收入取不同值时,考试成绩几乎不变)?

■ 在消费需求的经济分析中,经常假设价格增长1%会导致需求量降低一定的百分比。由于价格上涨1%所导致的需求量的百分比变化,被称为价格弹性(elasticity)。

回归模型设定为自然对数形式,我们可以应用该回归模型估计诸如上述的百分比关系。在介绍这些回归设定之前,我们复习一下指数函数和自然对数函数。

指数函数和自然对数。指数函数,与它的反函数——自然对数,在建立非线性函数模型中起着重要的作用。x 的**指数函数**(exponential function)是 e^x,即 e 以 x 次幂增长,这里 e 为常数2.71828…,指数函数也可写为 $\exp(x)$。**自然对数**(natural logarithm)是指数函数的反函数,即自然对数是 $x=\ln(e^x)$,或 $x=\ln[\exp(x)]$的函数。自然对数的底数是 e。尽管还有其他底数的对数,如底数为10,但本书中我们只考虑以 e 为底数的对数,即自然对数,所以当我们使用术语“对数”时,我们总是指“自然对数”。

对数函数 $y=\ln(x)$绘制在图6—4 中。注意,对数函数仅对 x 取正值时有定义。对数函数的斜率开始时陡峭,而后又变平了(尽管函数连续递增)。对数函数 $\ln(x)$的斜率是 $1/x$。

对数函数具有如下有用的性质:

$$\ln(1/x) = -\ln(x) \tag{6.12}$$

$$\ln(ax) = \ln(a) + \ln(x) \tag{6.13}$$

$$\ln(x/a) = \ln(x) - \ln(a) \tag{6.14}$$

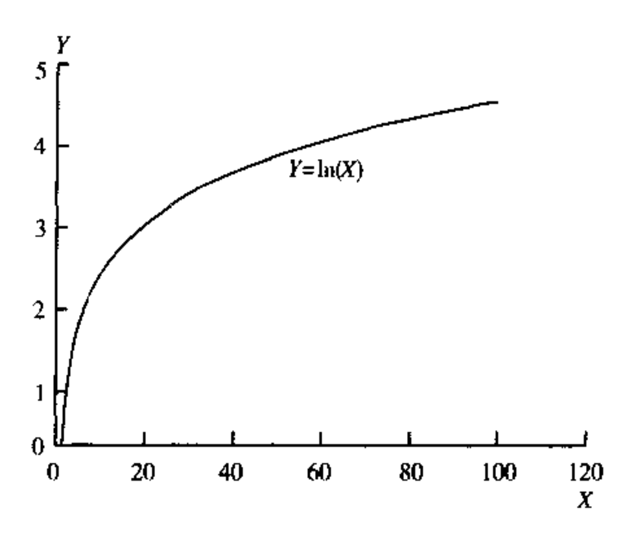
$$\ln(x^a) = a\ln(x) \tag{6.15}$$

对数和百分数。对数和百分数之间的联系依赖于一个关键的事实:当 Δx 很小时,$x+\Delta x$ 的对数与 x 的对数之差约等于 $\Delta x/x$,即 x 的百分比变化被100除,也就是说:

$$\ln(x+\Delta x) - \ln(x) \approx \frac{\Delta x}{x} \left(当\frac{\Delta x}{x}很小时\right) \tag{6.16}$$

这里“≈”意思是“约等于”。这个近似的严密推导要依靠微积分,但却很容易通过试用一些 x 和 Δx 的值加以证明。例如,当 $x=100$,$\Delta x=1$ 时,$\Delta x/x=1/100=0.01$(或1%),$\ln(x+\Delta x)-\ln(x)=\ln(101)-\ln(100)=0.00995$(或0.995%)。因此,$\Delta x/x$(值为0.01)非常接近于 $\ln(x+\Delta x)-\ln(x)$(值为0.00995)。当 $\Delta x=5$ 时,$\Delta x/x=5/100=0.05$,而 $\ln(x+\Delta x)-\ln(x)=\ln(105)-\ln(100)=0.04879$。





注：对数函数 $Y=\ln(X)$ 对于小的 X 值比对于大的 X 值陡峭，只对于 $X>0$ 时有定义，且有斜率 $1/X$。

图 6—4　对数函数 $Y=\ln(X)$

..个对数回归模型。有三种可能会用到对数的不同情况：X 通过取对数来进行转换，而 Y 不进行转换；Y 被变换成它的对数，而 X 不作转换；X 和 Y 都被转换成它们的对数。在每种情况下回归系数的解释是不同的。我们依次讨论这三种情况。

情况 1：X 取对数，而 Y 不取对数。在这种情况下，回归模型为：

$$Y_i=\beta_0+\beta_1\ln(X_i)+u_i,i=1,\cdots,n \tag{6.17}$$

由于 Y 没有取对数，而 X 取了对数，因此这种情况有时被称为线性对数模型（linear-log model）。

在线性对数模型中，X 变化 1%，Y 对应变化 $0.01\beta_1$。为了弄明白这一点，考虑总体回归函数在 X 变化 ΔX 的不同取值处之间的差异，这就是 $[\beta_0+\beta_1\ln(X+\Delta X)]-[\beta_0+\beta_1\ln(X)]=\beta_1[\ln(X+\Delta X)-\ln(X)]\approx\beta_1(\Delta X/X)$，这里最后一步使用了表达式(6.16)中的近似结果。如果 X 变化 1%，那么 $\Delta X/X=0.01$，因此，在这个模型中，X 变化 1% 与 Y 变化 $0.01\beta_1$ 相对应。

公式(6.17)中的回归模型与第 4 章的一元回归模型之间的惟一区别就是方程右边的变量现在是 X 的对数而不是 X 本身。为估计公式(6.17)中的系数 β_0 和 β_1，首先要计算新变量 $\ln(x)$，这用电子表格软件或统计软件很容易做到。然后，就可用 Y_i 对 $\ln(X_i)$ 的 OLS 回归来估计 β_0 和 β_1，可以用 t 统计量来检验关于 β_1 的假设，而 β_1 的 95% 的置信区间可被构造为 $\hat{\beta}_1\pm1.96SE(\hat{\beta}_1)$。

作为一个例子，让我们回到地区收入和考试成绩之间的关系上来。我们用公式(6.17)中的线性对数设定代替二次设定，用 OLS 估计这个回归得到：

$$\widehat{TestScore}=\underset{(3.8)}{557.8}+\underset{(1.40)}{36.42}\ln(Income),\bar{R}^2=0.561 \tag{6.18}$$

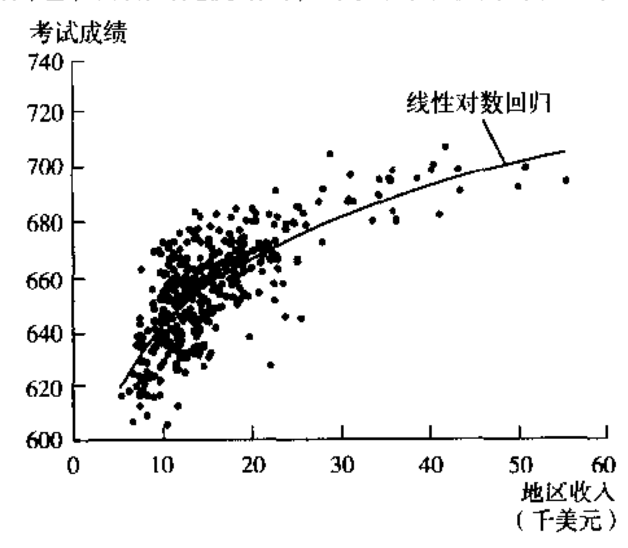
根据公式(6.18)，收入每增加 1%，考试成绩增加 $0.01\times36.42=0.36$ 分。

为了估计用初始单位千美元（而非对数）测度的 X 的变化对 Y 的效应，我们可以用重要概念 6.1 中的方法。例如，平均收入 10 000 美元的地区和平均收入 11 000 美元的地区之间考试成绩的预测差是多少？ΔY 的估计值是预测值之差：$\Delta\hat{Y}=[557.8+36.42\ln(11)]-[557.8+36.42\ln(10)]=36.42\times[\ln(11)-\ln(10)]=3.47$。同理，平均收入 40 000 美元的地区和平均收入 41 000 美元的地区之间的预测差为 $36.42\times[\ln(41)-\ln(40)]=0.90$。



因此，和二次设定一样，这个回归预测了在贫穷地区收入增加 1 000 美元对考试成绩的效应比在富裕地区大。

公式(6.18)中所估计的线性对数回归函数绘制在图 6—5 中。由于公式(6.18)中的回归因子是收入的自然对数而不是收入，因此所估计的回归函数不是一条直线。像图 6—3 中的二次回归函数一样，它在开始时是陡峭的，而随后对于较高的收入水平又变平了。



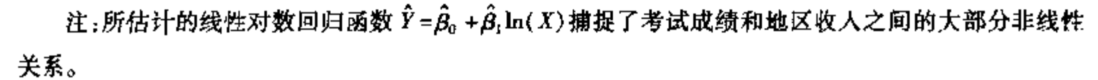

注：所估计的线性对数回归函数 $\hat{Y}=\hat{\beta}_0+\hat{\beta}_1\ln(X)$ 捕捉了考试成绩和地区收入之间的大部分非线性关系。

图 6—5　线性对数回归函数



情况 2：Y 取对数，而 X 不取对数。在这种情况下，回归模型为：

$$\ln(Y_i)=\beta_0+\beta_1 X_i+u_i \tag{6.19}$$

由于 Y 取对数，而 X 不取对数，因此这种情况被称为对数线性模型(log-linear model)。

在对数线性模型中，X 每变化 1 个单位($\Delta X=1$)，Y 对应变化 $100\times\beta_1\%$。为了弄明白这一点，比较 X 变化 ΔX 的不同取值处的 $\ln(Y)$ 的期望值。给定 X 条件下 $\ln(Y)$ 的期望值为 $\ln(Y)=\beta_0+\beta_1 X$。当 X 为 $X+\Delta X$ 时，由等式 $\ln(Y+\Delta Y)=\beta_0+\beta_1(X+\Delta X)$ 给出期望值。因此，这两个期望值之差为 $\ln(Y+\Delta Y)-\ln(Y)=[\beta_0+\beta_1(X+\Delta X)]-(\beta_0+\beta_1 X)=\beta_1\Delta X$。然而，根据公式(6.16)中的近似，如果 $\beta_1\Delta X$ 很小，那么 $\ln(Y+\Delta Y)-\ln(Y)\approx\Delta Y/Y$，这样，$\Delta Y/Y\approx\beta_1\Delta X$。如果 $\Delta X=1$ 使得 X 改变 1 个单位，那么 $\Delta Y/Y$ 改变 β_1。变换为百分数，X 变化 1 个单位与 Y 变化 $100\times\beta_1\%$ 相对应。

举例说明，我们回到 3.6 节的实证例子，即大学毕业生的年龄和收入之间关系的问题上。很多雇佣合同中明确规定，每多服务一年，工人会得到在其工资基础上的一定百分比的增加。这个百分比关系表明可以应用估计公式(6.19)中的对数线性设定，即从总体平均来看，年龄(X)每增加一岁是与收入(Y)的某一恒定百分比的增加相联系的。首先要计算新的因变量 $\ln(Earnings_i)$，然后未知系数 β_0 和 β_1 可以用 $\ln(Earnings_i)$ 对 Age_i 回归的 OLS 方法进行估计。当使用 1999 年当前人口调查(附录 3.1 中所描述的数据)中 12 077 个大学毕业生的观察值来估计这个关系时，得到如下的估计方程：

$$\widehat{\ln(Earings)}=\underset{(0.024)}{2.453}+\underset{(0.0006)}{0.0128}\ Age, \bar{R}^2=0.0387 \tag{6.20}$$

根据这个回归，年龄每增加一岁，收入预计增长 1.28%($0.0128\times100\%$)。



双对数回归的 $\overline{R}^2(0.577)$ 比对数线性回归的 $\overline{R}^2(0.497)$ 高这一点相一致。即使如此，双对数设定对数据的拟合也不是很好：在较低的收入值处，大部分观察值落在双对数回归曲线的下方，而在收入的中间范围内，大部分观察值落在所估计的回归函数的上方。

三个对数回归模型在重要概念 6.2 中总结。

在比较对数设定时的一个困难。哪一个对数回归模型拟合数据最好？正如我们在公式(6.23)和公式(6.24)中所看到的，$\overline{R}^2$ 可以用来比较对数线性模型和双对数模型，这里双对数模型恰巧有较高的 $\overline{R}^2$。同理，$\overline{R}^2$ 可以用来比较公式(6.18)的线性对数回归和 Y 对 X 的线性回归。在考试成绩和收入的回归中，线性对数回归模型的 $\overline{R}^2$ 为 0.561，而线性回归的 $\overline{R}^2$ 为 0.508，所以，线性对数模型拟合数据的效果更好。

我们如何比较线性对数模型和双对数模型呢？$\overline{R}^2$ 不能用来比较这两个回归，因为它们的因变量是不同的（一个是 Y_i，另一个是 $\ln(Y_i)$）。回想一下，$\overline{R}^2$ 测度的是因变量的方差中被回归因子所解释的部分。由于双对数模型和线性对数模型中的因变量是不同的，因此比较它们的 $\overline{R}^2$ 是没有意义的。

由于这个问题的存在，在一个特定的应用中，最好的方法就是用经济理论和你自己的认识或其他专家对这个问题的认识来判定 Y 取对数是否有意义。例如，劳动经济学家通常用对数来建立收入模型，因为工资的比较、合同工资的增加等等都最自然地常以百分比形式来讨论。在建立考试成绩模型时，以考试的成绩分数而不是考试成绩分数的百分比讨论考试结果，看上去是很自然的（至少对于我们而言），所以我们集中讨论因变量为考试成绩分数而不是它的对数的模型。

重要概念 6.2

对数回归：三种情况

对数可以用来转换因变量 Y、自变量 X 或同时转换 Y 与 X（但是它们必须都是正数）。表 6—1 归纳了这三种情况以及回归系数 β_1 的解释。在每种情况下，β_1 可以通过在对因变量和/或自变量取对数后，使用 OLS 进行估计。

表 6—1　　对数回归：三种情况

情况	回归设定	β_1 的解释
1	$Y_i=\beta_0+\beta_1\ln(X_i)+u_i$	X 变化 1%，Y 对应变化 $0.01\beta_1$
2	$\ln(Y_i)=\beta_0+\beta_1X_i+u_i$	X 变化 1 个单位($\Delta X=1$)，Y 对应变化 $100\beta_1\%$
3	$\ln(Y_i)=\beta_0+\beta_1\ln(X_i)+u_i$	X 变化 1%，Y 对应变化 $\beta_1\%$，因此 β_1 是 Y 关于 X 的弹性

Y 取对数时计算 Y 的预测值①。如果因变量 Y 已进行对数变换，那么所估计出的回归可以直接用来计算 $\ln(Y)$ 的预测值。不过，计算 Y 本身的预测值则有一些小的麻烦。

为了弄明白这一点，考虑公式(6.19)中的对数线性回归模型，并将它重写，重写后的方程以 Y 来设定而不是以 $\ln(Y)$ 来设定。为此，对公式(6.19)的两边都取指数，结果是：

$$Y_i=\exp(\beta_0+\beta_1X_i+u_i)=e^{\beta_0+\beta_1X_i}e^{u_i} \tag{6.25}$$

如果 u_i 独立分布于 X_i，那么给定 X_i 条件下 Y_i 的期望值为 $E(Y_i|X_i)=E(e^{\beta_0+\beta_1X_i}e^{u_i}|X_i)=e^{\beta_0+\beta_1X_i}E(e^{u_i})$。问题是，即使 $E(u_i)=0$，也有 $E(e^{u_i})\neq 1$。这样，不能简单地通过取 $\hat{\beta}_0+\hat{\beta}_1X_i$ 的指数函数，即通过设 $\hat{Y}_i=e^{\hat{\beta}_0+\hat{\beta}_1X_i}$ 得到 Y_i 的合适的预测值。由于缺少因子 $E(e^{u_i})$，因此这个预测值是有偏的。

① 这一部分内容更高深一些，跳过这一部分不会影响理解上的连续性。



这个问题的解决办法就是估计因子 $E(e^{u_i})$，并且在计算 Y 的预测值时使用这个估计值，但是这会变得很复杂，我们就不再进一步探讨了。

另一种"解决方法"，也是本书中所采取的方法，就是计算 Y 的对数的预测值，而不把它们转化为原来的单位。实际上，通常这是可接受的，因为当因变量被设定为对数时，整个分析中只用对数设定（和相关的百分比解释）通常是最自然的。

6.2.3 考试成绩和地区收入的多项式模型和对数模型



实际上，经济理论或专家的判断可能会提出一个函数形式来使用，但是最终总体回归函数的真实形式却是未知的。因此，实际上拟合非线性函数必须判定哪一种方法或方法的组合效果最好。作为一个例子，我们来比较地区收入和考试成绩之间关系的对数模型和多项式模型。

多项式设定。我们考虑用 *Income* 的二次幂（公式(6.2)）和三次幂（公式(6.11)）所设定的两个多项式模型。因为公式(6.11)中 $Income^3$ 的系数在5%的水平下是显著的，三次设定比二次设定有所改进，所以我们选择三次模型作为首选的多项式设定。

对数设定。公式(6.18)中的对数设定看上去为数据提供了很好的拟合，但是我们没有对其进行正式检验。一种检验方法就是用收入对数的更高次幂来扩展这个模型。如果增加的项在统计上并不是异于0的，那么我们可以得出结论：相对于对数多项式函数，公式(6.18)中的设定不能被拒绝，在这个意义上说它是充分的。因此，所估计的三次回归（用收入对数的幂来设定）为：

$$\widehat{TestScore} = \underset{(79.4)}{486.1} + \underset{(87.9)}{113.4}\ln(Income) - \underset{(31.7)}{26.9}[\ln(Income)]^2 + \underset{(3.74)}{3.06}[\ln(Income)]^3, \bar{R}^2 = 0.560 \tag{6.26}$$

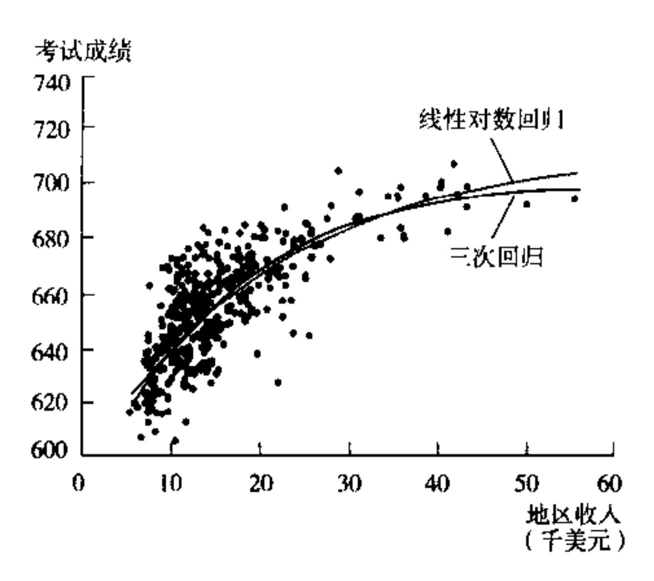
三次项系数的 t 统计量为0.818，因此真实系数为零的零假设在10%的显著性水平下不能被拒绝。检验二次项和三次项系数均为零的联合假设的 F 统计量为0.44，p 值为0.64，所以这个联合零假设在10%的显著性水平下不能被拒绝。因而，公式(6.26)中的三次对数模型并没有比公式(6.18)中的收入对数的线性模型提供一个统计上更显著的改进。

比较三次设定和线性对数设定。图6—7绘制了来自公式(6.11)的三次设定和公式(6.18)的线性对数设定所估计的回归函数的图形。这两个估计的回归函数是非常相似的。比较这两个设定的一个统计工具是 $\bar{R}^2$。对数回归的 $\bar{R}^2$ 为0.561，而三次回归的 $\bar{R}^2$ 为0.555。由于根据 $\bar{R}^2$ 大小判断，对数设定稍有优势，并且由于这个设定不需要用收入对数的较高次多项式拟合这些数据，因此我们采用公式(6.18)中的对数设定。

6.3 自变量之间的交互作用

在本章的引言中，我们想知道在很多学生仍在学习英语的地区，降低学生—教师比对考试成绩的影响是否比在很少学生仍在学习英语的地区更大。例如，如果仍在学习英语的学生个别地受益于一对一的或小范围的教学指导，这种情况可能就会出现。如果是这样的话，一个地区有很多的英语学习者，将会与学生—教师比的变化按以下的方式发生交互作用，即学生—教师比的变化对考试成绩的影响依赖于学生中英语学习者的比率。

本节解释如何将两个自变量之间的这种交互作用体现到多元回归模型中。学生—教师比与英语学习者的比率之间可能的交互作用，就是说明一个自变量的变化对 Y 的效应依赖



注:在这个样本中,所估计的三次回归函数(公式(6.11))和所估计的线性对数回归函数(公式(6.11))几乎是相同的。

图 6—7 线性对数回归函数和三次回归函数



于另一个自变量的值的一个一般性的例子。我们考虑三种情况:两个自变量都是二元的;一个是二元的,而另一个是连续的;两个变量都是连续的。

6.3.1 两个二元变量之间的交互作用

考虑对数收入(Y_i,这里 $Y_i=\ln(Earnings_i)$)对两个二元变量——个人的性别(D_{1i},如果第 i 个人为女性,那么 $D_{1i}=1$)和他(她)是否有大学学历(D_{2i},如果第 i 个人大学毕业,那么 $D_{2i}=1$)的总体回归。Y_i 对这两个二元变量的总体线性回归是:

$$Y_i=\beta_0+\beta_1 D_{1i}+\beta_2 D_{2i}+u_i \tag{6.27}$$

在这个回归模型中,β_1 是在保持教育年限不变的情况下,女性对对数收入的效应,β_2 是在保持性别不变的情况下,有大学学历对对数收入的效应。

公式(6.27)中的设定有个重要的缺陷:在保持性别不变的情况下,这个设定中具有大学学历的效应对于男性和女性而言都是相同的。然而,没有理由认为一定会这样。用数学语言表述,在 D_{1i}保持不变的情况下,D_{2i}对 Y_i 的效应可能会依赖于 D_{1i}的值。换句话说,在性别与有大学学历之间可能存在交互作用,使得学历在人才市场上的价值对于男性和女性而言是不同的。

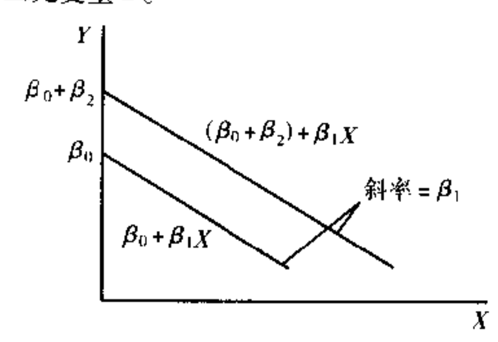
尽管公式(6.27)中的设定没有考虑性别与获取大学学历之间的这种交互作用,但很容易通过修正方程的设定使方程考虑这种交互作用,即通过引入另一个回归因子——两个二元变量之积 $D_{1i}\times D_{2i}$的方式来实现,其相应的回归方程是:

$$Y_i=\beta_0+\beta_1 D_{1i}+\beta_2 D_{2i}+\beta_2(D_{1i}\times D_{2i})+u_i \tag{6.28}$$

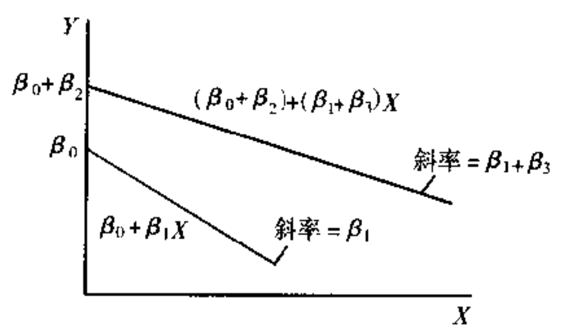
上式中新的回归因子,即乘积 $D_{1i}\times D_{2i}$,被称为**交互作用项**(interaction term),或**交互作用回归因子**(interaction regressor),而公式(6.28)中的总体回归模型被称为二元变量的交互作用回归模型(interaction regression model)。

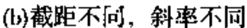
公式(6.28)中的交互作用项允许具有大学学历(将 D_{2i}从 $D_{2i}=0$ 变化到 $D_{2i}=1$)对对数收入(Y_i)的总体效应依赖于性别(D_{1i})的变化。为了在数学上证明这一点,利用重要概念 6.1 中所给出的一般方法来计算 D_{2i}的变化的总体效应。第一步,在给定 D_{1i}的值的情况下,

二元变量 D。

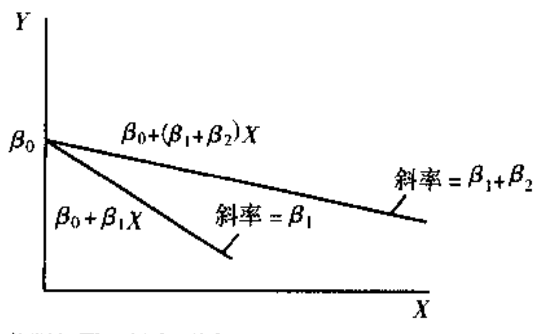


(a)截距不同，斜率相同





(b)截距不同，斜率不同



(c)截距相同，斜率不同

注：二元变量和连续变量的交互作用可能会得到三种不同的总体回归函数：

(a) $\beta_0+\beta_1X+\beta_2D$ 考虑了不同的截距，但具有相同的斜率；

(b) $\beta_0+\beta_1X+\beta_2D+\beta_3(X\times D)$ 考虑了不同的截距和不同的斜率；

(c) $\beta_0+\beta_1X+\beta_2(X\times D)$ 考虑了相同的截距，但具有不同的斜率。

图 6—8　使用二元变量和连续变量的回归函数

在图 6—8(a)中，两条回归直线的不同之处只是它们的截距不同。相应的总体回归模型是：

$$Y_i=\beta_0+\beta_1X_i+\beta_2D_i+u_i \tag{6.31}$$

这就是我们熟悉的多元回归模型，总体回归函数是关于 X_i 和 D_i 线性的函数。当 $D_i=0$ 时，总体回归函数是 $\beta_0+\beta_1X_i$，因此，截距是 β_0，斜率是 β_1。当 $D_i=1$ 时，总体回归函数是 $\beta_0+\beta_1X_i+\beta_2$，因此，斜率仍是 β_1，但截距是 $\beta_0+\beta_2$。所以，如图 6—8(a)所示，β_2 是两条回归线的截距之差。根据收入的例子所述，β_1 是在保持大学学历状态不变的情况下，工龄每增加一年对对数收入的效应；而 β_2 是在保持工龄不变的情况下，大学学历对对数收入的效应。在这个设定中，工龄增加一年的效应对大学毕业生和非大学毕业生都是一样的，也就是说，图 6—8(a)中的两条直线具有相同的斜率。

在图 6—8(b)中，两条直线具有不同的斜率和截距。不同的斜率使得工龄增加一年的效应对大学毕业生和非大学毕业生而言有所差异。为了考虑不同的斜率，添加一个交互作用项到公式(6.31)中：

$$Y_i=\beta_0+\beta_1X_i+\beta_2D_i+\beta_3(X_i\times D_i)+u_i \tag{6.32}$$

其中，$X_i\times D_i$ 是个新的变量。为了解释这个回归的系数，应用重要概念 6.3 中的方法。应用该方法我们会发现，如果 $D_i=0$，那么总体回归函数是 $\beta_0+\beta_1X_i$，但如果 $D_i=1$，那么总体回归函数是 $(\beta_0+\beta_2)+(\beta_1+\beta_3)X_i$。因此，如图 6—8(b)所示，这个设定考虑到了联系 Y_i 与 X_i(依赖于 D_i 的值)的两个不同的总体回归函数。这两个设定的截距之差是 β_2，斜率之



和截距都相同的联合检验，但是使用 t 统计量对单个假设的检验却没有拒绝。出现这种情况的原因是回归因子 $HiEL$ 和 $STR \times HiEL$ 之间是高度相关的。这导致单个系数有很大的标准误。即使这样，要断定哪个系数是非零的也是不可能的，但也有相当强的证据拒绝系数均为零的假设。

第四，检验“学生—教师比不进入这个模型设定”的假设，可以通过计算 STR 和交互项的系数都为零的联合假设的 F 统计量来完成。这个 F 统计量是 5.64，p 值为 0.004。因而，学生—教师比的系数在 1% 的显著性水平下在统计上是显著的。

重要概念 6.4

二元变量和连续变量之间的交互作用

通过使用交互作用项 $X_i \times D_i$，联系 Y_i 和连续变量 X_i 之间的总体回归直线可能会有个依赖于二元变量 D_i 的斜率。这有三种可能情况：

1. 截距不同，斜率相同（见图 6—8(a)）：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + u_i$$

2. 截距和斜率都不同（见图 6—8(b)）：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 D_i + \beta_3 (X_i \times D_i) + u_i$$

3. 截距相同，斜率不同（见图 6—8(c)）：

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 (X_i \times D_i) + u_i$$

6.3.3 两个连续变量之间的交互作用

现在假设两个自变量（X_{1i}和 X_{2i}）都是连续的。一个例子是，Y_i 表示第 i 个工人的对数收入，X_{1i}表示他（或她）的工龄，而 X_{2i}表示他（或她）的上学年数。如果总体回归函数是线性的，那么工龄增加一年对工资的效应不依赖于受教育的年数，或者说，增加一年的教育对工资的效应不依赖于工龄。然而，实际上这两个变量之间可能存在交互作用，使得工龄增加一年对工资的效应依赖于受教育的年数。这种交互作用可以通过用交互作用项 X_{1i}与 X_{2i}之积扩展线性回归模型来建模：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \beta_3 (X_{1i} \times X_{2i}) + u_i \tag{6.35}$$

交互作用项允许 X_1的单位变化的效应依赖于 X_2。为了弄明白这一点，应用重要概念 6.1 中计算非线性回归模型效应的一般方法。公式(6.35)中的交互作用回归函数所计算的公式(6.6)中的差分为 $\Delta Y = (\beta_1 + \beta_3 X_2)\Delta X_1$（见练习 6.5(a)）。因此，在保持 X_2 不变的情况下，X_1 的变化对 Y 的效应为：

$$\frac{\Delta Y}{\Delta X_1} = \beta_1 + \beta_3 X_2 \tag{6.36}$$

它依赖于 X_2。例如，在上述收入的例子中，如果 β_3 是正的，那么工人所受到的教育每增加一年，工龄增加一年对对数收入的效应会以数量 β_3 增加。

一般兴趣框



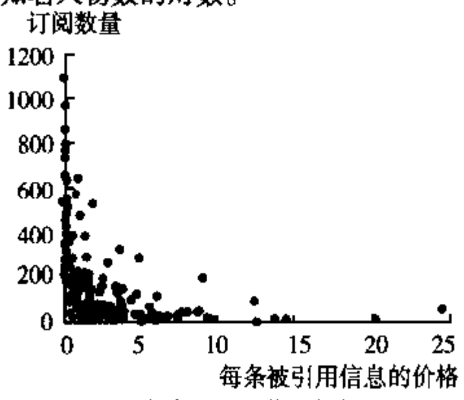
对经济学杂志的需求

专业经济学家们跟踪其专业领域的最新研究。经济学中的大多数近期研究成果首先刊登在经济学杂志上，所以经济学家们或他们的图书馆都订阅经济学杂志。

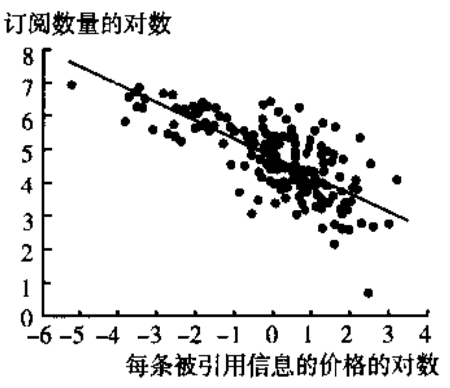
图书馆对经济学杂志的需求弹性有多大呢？为了求出这个弹性，我们使用 2000 年 180

种经济学杂志的数据，分析了美国图书馆订阅杂志的数量(Y_i)与订阅价格之间的关系。因为杂志产品并不只是印刷过的纸张而已，而是杂志所包含的思想，所以杂志的价格当然不能用每年的价值或每页的价值测度，而应该用每个思想的价值来测度。虽然我们不能直接测度"思想"，但是一个比较好的间接测度就是一本杂志中的论文随后被其他研究人员所引用的次数。因此，我们把杂志中"每条被引用信息的价格"作为该杂志的价格测度。这个价格范围非常大，每条引用从 0.5 美分(《美国经济评论》)到 20 美分或更多。一些杂志的每条引用价格是很昂贵的，因为它们很少被引用，另外一些杂志则因为每年图书馆订阅价格很高而变得很贵：2000 年，一个图书馆订阅《经济计量学杂志》大约要花费 1 900 美元，是订阅《美国经济评论》杂志价格的 40 倍!

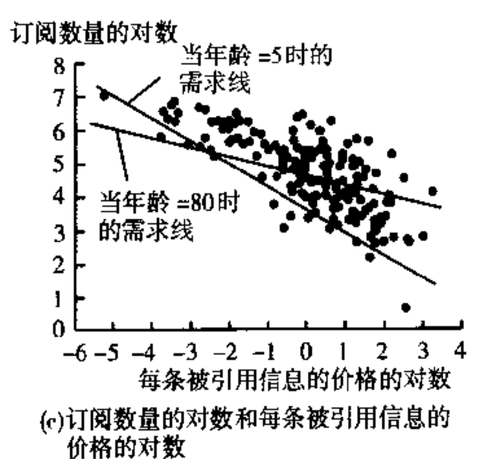
因为我们感兴趣的是估计其弹性，所以我们使用双对数设定(见重要概念 6.2)。图 6—9(a)和图 6—9(b)中的散点图为这种转换提供了经验支持。因为一些年代最久的和最有声望的杂志每条被引用信息的价格是最便宜的，所以对数需求量对对数价格的回归可能会存在遗漏变量偏差。因此我们的回归中包括两个控制变量，杂志年龄的对数和每年该杂志上知名人物数的对数。



(a)订阅数量和每条被引用信息的价格



(b)订阅数量的对数和每条被引用信息的价格的对数



(c)订阅数量的对数和每条被引用信息的价格的对数

注：对于 2000 年的 180 家经济学杂志而言，如图 6—9(a)所示，在美国图书馆订阅数量(需求量)和每条被引用信息的图书馆价格(价格)之间存在一个非线性的反比例关系。但如图 6—9(b)所示，对数需求量和对数价格之间的关系看上去是近似线性的。图 6—9(c)表明了对新($Age=5$)杂志的需求比对老($Age=80$)杂志的需求更富弹性。

图 6—9 经济学杂志的图书馆订阅数量和价格

表 6—2 经济学杂志需求的估计量

因变量:2000 年美国图书馆的订阅数量的对数;180 个观测值				
回归因子	(1)	(2)	(3)	(4)
ln(*Price per citation*)	-0.533^{**}	-0.408^{**}	-0.961^{**}	-0.899^{**}
	(0.034)	(0.044)	(0.160)	(0.145)
[ln(*Price per citation*)]^2			0.017	
			(0.025)	
[ln(*Price per citation*)]^3			0.0037	
			(0.0055)	
ln(*Age*)		0.424^{**}	0.373^{**}	0.374^{**}
		(0.119)	(0.118)	(0.118)
ln(*Age*) × ln(*Price per citation*)			0.156^{**}	0.141^{**}
			(0.052)	(0.040)
ln(*Characters* ÷ 1 000 000)		0.206^{*}	0.235^{*}	0.229^{*}
		(0.098)	(0.098)	(0.096)
	4.77^{**}	3.21^{**}	3.41^{**}	3.43^{**}
截距项	(0.055)	(0.38)	(0.38)	(0.38)
F 统计量和总括性统计量				
检验二次项和三次项系数的			0.25	
F 统计量(*p* 值)			(0.779)	
SER	0.750	0.705	0.691	0.688
$\bar{R}^2$	0.555	0.607	0.622	0.626

注:*F* 统计量检验[ln(*Price per citation*)]^2 和[ln(*Price per citation*)]^3 的系数都为零的假设。在系数下方的括号中给出了标准误,*F* 统计量下方的括号中给出了 *p* 值。单个系数在 *5% 的水平或是 **1% 的水平下是显著的。

回归结果总结在表 6—2 中。表中的结果使我们得出如下结论(看看你是否能够在表中找到这些结论的根据):

1. 对较老杂志的需求弹性比对较新杂志的需求弹性小;
2. 表 6—2 中的证据支持价格对数的线性函数,而不是价格对数的三次函数;
3. 在价格和杂志年龄保持不变的情况下,具有更多知名人物的杂志需求量更大。

那么,对经济学杂志的需求弹性是多少呢? 它依赖于杂志的年龄。有 80 年历史的老杂志和有 5 年历史的新杂志的需求曲线被添加到图 6—9(c) 的散点图中;对较老杂志的需求弹性是 -0.28($SE=0.06$),而对较新杂志的需求弹性是 -0.67($SE=0.08$)。

对经济学杂志的需求非常缺乏弹性:需求对价格非常不敏感,尤其对较老的杂志来说。对 图书馆而言,收藏手边最新的研究是必需品,而不是奢侈品。通过比较,专家估计烟的需



求弹性是在 -0.3 到 -0.5 之间。经济学杂志看上去和香烟一样让人上瘾,但它对健康更有益![1]

在 X_1 保持不变时,相似的计算表明了 X_2 变化 ΔX_2 对 Y 的效应是$\frac{\Delta Y}{\Delta X_2}=(\beta_2+\beta_3X_1)$。

将这两个效应放在一起,表明交叉项的系数 β_3 是 X_1 和 X_2 都增加 1 个单位的效应,它比只增加 1 个单位的 X_1 和只增加 1 个单位的 X_2 的效应之和还要大。也就是说,如果 X_1 变化 ΔX_1,X_2 变化 ΔX_2,那么 Y 的期望变化为 $\Delta Y=(\beta_1+\beta_3X_2)\Delta X_1+(\beta_2+\beta_3X_1)\Delta X_2+\beta_3\Delta X_1\Delta X_2$(练习 6.5(c))。第一项是在 X_2 保持不变的情况下,变化 X_1 的效应;第二项是在 X_1 保持不变的情况下,变化 X_2 的效应;而最后一项 $\beta_3\Delta X_1\Delta X_2$ 是变化 X_1 和 X_2 的额外效应。

两个变量之间的交互作用在重要概念 6.5 中总结。

当交互作用与对数转换组合在一起时,它们可被用来估计价格弹性,这时价格弹性依赖于商品的特征(见本章一般兴趣框中的例子)。

在学生—教师比和英语学习者的百分比案例中的应用。前面的例子中考虑了学生—教师比和一个表示英语学习者的百分比大小的二元变量之间的交互作用。研究这个交互作用的另一种不同的方法,是检验学生—教师比与英语学习者的百分比($PctEL$)这个连续变量之间的交互作用。所估计的交互回归方程为:

$$\widehat{TestScore}=\underset{(11.8)}{686.3}-\underset{(0.59)}{1.12}\,STR-\underset{(0.37)}{0.67}\,PctEL+\underset{(0.019)}{0.0012}\,(STR\times PctEL),\ \bar{R}^2=0.422 \qquad (6.37)$$

当英语学习者的百分比取其中位数($PctEL=8.85$)时,考试成绩和学生—教师比之间关系的估计直线方程的斜率为 -1.11(-1.12+0.0012×8.85)。当英语学习者的百分比取其第 75 位百分位数($PctEL=23.0$)时,这条估计的直线会变得更平坦,斜率为 -1.09(-1.12+0.0012×23.0)。也就是说,对一个英语学习者的百分比为 8.85% 的地区而言,学生—教师比减少 1 个单位的估计效应是使考试成绩增加 1.11 分,但是对一个英语学习者的百分比为 23.0% 的地区而言,学生—教师比减少 1 个单位预计只会使考试成绩增加 1.09 分。可是,这些估计的效应之差在统计上并不显著:检验交叉项的系数是否为零的 t 统计量为 $t=0.0012/0.019=0.06$,它在 10% 的水平下不显著。

重要概念 6.5

多元回归中的交互作用

两个自变量 X_1 和 X_2 之间的交叉项是它们之积 $X_1\times X_2$。方程中包括这个交叉项,允许 X_1 的变化对 Y 的效应依赖于 X_2 的值,反之,允许 X_2 的变化对 Y 的效应依赖于 X_1 的值。

$X_1\times X_2$ 的系数是 X_1 和 X_2 每增加 1 个单位的效应,它比只改变 1 个单位的 X_1 和只改变 1 个单位的 X_2 的效应之和大。无论 X_1 和/或 X_2 是连续变量还是二元变量,这个结论都成立。

为了使讨论集中在非线性模型上,我们在 6.1 节 ~6.3 节的设定讨论中排除了诸如学生经济背景等额外的控制变量,因此,可以证明这些结果受遗漏变量偏差的影响。为了得到关于降低学生—教师比对考试成绩的效应的实质性的结论,必须用控制变量来扩展这些非线性设定,这就是我们现在所要转入的一个练习。

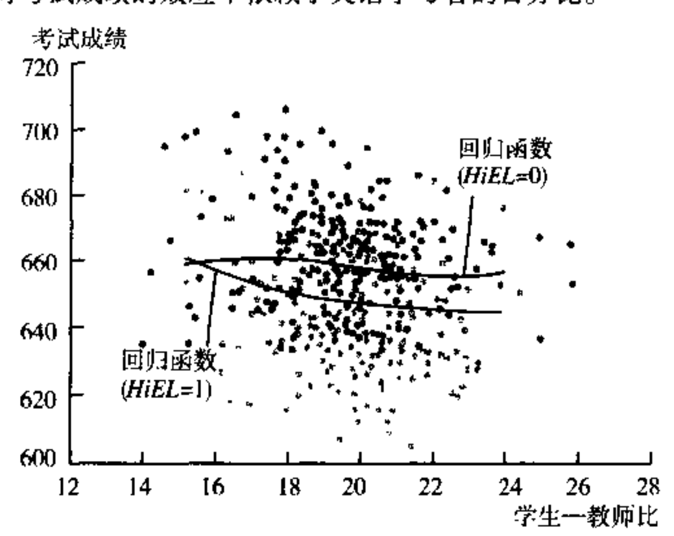
[1] 这些数据是由圣巴巴拉市加利福尼亚州立大学经济系的 Theodore Bergstrom 教授热心提供的。如果你对了解更多的经济学杂志上的经济学问题感兴趣,参见 Bergstrom(2001)。

表 6—3　　**考试成绩的非线性回归模型**

因变量：地区内的平均考试成绩；420 个观察值							
回归因子	(1)	(2)	(3)	(4)	(5)	(6)	(7)
学生—教师比(*STR*)	-1.00**	-0.73**	-0.97	-0.53	64.33**	83.70**	65.29**
	(0.27)	(0.26)	(0.59)	(0.34)	(24.86)	(28.50)	(25.26)
STR^2					-3.42**	-4.38**	-3.47**
					(1.25)	(1.44)	(1.27)
STR^3					0.059**	-0.075**	0.060**
					(0.021)	(0.024)	(0.021)
英语学习者的百分比	0.122**	-0.176**					-0.166**
	(0.033)	(0.034)					(0.034)
英语学习者的百分比≥10%？			5.64	5.50	-5.47**	816.1	
(二元变量，*HiEL*)			(19.51)	(9.80)	(1.03)	(327.7)	
$HiEL \times STR$			-1.28	-0.58		-123.3*	
			(0.97)	(0.50)		(50.2)	
$HiEL \times STR^2$						6.12*	
						(2.54)	
$HiEL \times STR^3$						-0.101*	
						(0.043)	
享有午餐补助	-0.547**	-0.398**		-0.411**	-0.420**	-0.418**	-0.402**
的学生的百分比	(0.024)	(0.033)		(0.029)	(0.029)	(0.029)	(0.033)
地区平均收入		11.57**		12.12**	11.75**	11.80**	11.51**
(对数)		(1.81)		(1.80)	(1.78)	(1.78)	(1.81)
截距	700.1**	658.6**	682.2**	653.6**	252.0	122.3	244.8
	(5.6)	(8.6)	(11.9)	(9.9)	(163.9)	(185.5)	(165.7)
联合假设的 *F* 统计量和 *p* 值							
(a) 所有 *STR* 变量			5.64	5.92	6.31	4.96	5.91
和交叉项 =0			(0.004)	(0.003)	(<0.001)	(<0.001)	(0.001)
(b) $STR^2, STR^3 = 0$					6.17	5.81	5.96
					(<0.001)	(0.003)	(0.003)
(c) $HiEL \times STR, HiEL \times STR^2$						2.69	
$HiEL \times STR^3 = 0$						(0.046)	
SER	9.08	8.64	15.88	8.63	8.56	8.55	8.57
$\bar{R}^2$	0.773	0.794	0.305	0.795	0.798	0.799	0.798

注：这些回归是用附录 4.1 中所描述的加利福尼亚州 K—8 校区的数据来估计的。在系数下面的括号给出了标准误，在 *F* 统计量下面的括号给了 *p* 值。单个系数在 * 5% 或 ** 1% 的显著性水平下在统计上是显著的。

之间的差异只反映了这些具有很低的学生—教师比的极少数地区的差异。因此，根据图6—11 我们得出结论：对于我们所拥有的数据所反映的学生—教师比的数值范围来说，学生—教师比的变化对考试成绩的效应不依赖于英语学习者的百分比。



注：低英语学习者百分比地区（*HiEL*=0）用灰色的点表示，而 *HiEL*=1 的地区用彩色的点表示。对于 17≤*STR*≤23 而言，来自表 6—3 中第(6)列的 *HiEL*=1 的三次回归函数大约比 *HiEL*=0 的三次回归函数低 10 分，但在其他情况下，这两个函数在这个范围内有相似的形状和斜率。对于很大和很小的 *STR* 值而言，回归函数的斜率差异很大，但这里只有极少数的观察值。

图 6—11 英语学习者百分比高低不同的地区的回归函数

6.4.2 结论总结

我们所讨论的这些结果可以让我们回答本节开始所提出的三个问题。

第一，在控制了经济背景变量之后，该地区英语学习者百分比的高低对学生—教师比的变化对考试成绩的效应没有显著的影响。在线性模型设定中，不存在显著的统计证据表明这种差异存在。回归(6)中的三次设定（在 5% 的显著性水平下）提供了对于英语学习者百分比高低不同的地区回归函数是不同的这一结论在统计上的显著证据。然而，如图 6—11 所示，在包含大部分样本数据的学生—教师比的数值范围内，所估计的回归函数具有相似的斜率。

第二，在控制了经济背景变量之后，存在学生—教师比对考试成绩的非线性效应的证据。这个影响效应在 1% 的水平下在统计上是显著的（STR^2 和 STR^3 的系数在 1% 的水平下总是显著的）。

第三，我们现在可以回到第 4 章所提出的教育主管所关心的问题。她想知道降低学生—教师比 2 个单位对考试成绩的效应。在线性设定(2)中，这个效应不依赖于学生—教师比本身，由于这个降低所得出的估计效应是使考试成绩提高 1.46（(−0.73)×(−2)）分。在非线性设定中，这个效应依赖于学生—教师比的值。如果目前她所在地区的学生—教师比为 20，而她正打算将其削减到 18，那么根据回归(5)，这一削减的估计效应是使考试成绩提高 3.00 分，而根据回归(7)，这个估计效应是 2.93。如果目前她所在地区的学生—教师比为 22，而她打算将其削减到 20，那么根据回归(5)，这一削减的估计效应是使考试成绩提高 1.93 分，而根据回归(7)，这个估计效应是 1.90。非线性设定的估计效应表明，如果学生—教师比已经很小，那么削减学生—教师比的效应会更大一些。



6.5 结论

本章提出了几种建立非线性回归函数模型的方法。由于这些模型是多元回归模型的变形，所以未知系数可以用 OLS 进行估计，而关于它们的值的假设可以用第 5 章中所描述的 t 统计量和 F 统计量进行检验。在这些模型中，在保持其他自变量 $X_2,\cdots,X_k$ 不变的情况下，一个自变量 X_1 的变化对 Y 的期望效应一般依赖于 $X_1,X_2,\cdots,X_k$ 的值。

本章中有很多不同的模型，而在一个给定的应用中，你不可能会因对该用哪一个模型感到为难而受到责备，那么在实际中你该如何分析可能的非线性关系呢？6.1 节列出了一种一般性的分析方法，但是这种方法要求你要沿着它一路进行决策和判断。如果有一种在每个应用中你都能够遵循的总是有效的方法，那么将会非常方便，但是在实际中数据分析很少会如此简单。

在设定非线性回归函数中惟一最重要的步骤是"使用你的脑子"。在观察数据之前，根据经济理论或者专家的判断，你能想出为什么总体回归函数的斜率可能会依赖于这个或那个自变量值的原因吗？如果你能想出，那么你预期会有什么样的依赖性呢？而且最重要的是，什么样的非线性关系（如果存在的话）可能会对描述你所提出的实质性问题有较大的意义？对这些问题的仔细回答会使你的分析越来越接近目标。例如，在考试成绩这一案例的应用中，这种推理会使我们进一步研究，雇佣更多的教师是否会对仍在学习英语的学生的百分比较大的地区有更大的影响，因为这些学生可能会从更多的个人关注中得到不同的益处。通过使问题变得精确，我们能够得到一个精确的答案：在控制了学生的经济背景之后，我们发现没有统计上的显著证据证明存在这种交互作用。

总结

1. 在非线性回归中，总体回归函数的斜率依赖于一个或多个自变量的值。
2. 自变量的变化对 Y 的影响，可以通过估计自变量在两个值处的回归函数来计算。这个方法在重要概念 6.1 中总结。
3. 一个多项式回归包含 X 的各次幂作为回归因子。二次回归包含 X 和 X^2，而三次回归包含 X,X^2 和 X^3。
4. 对数的一个小的变化，可被解释为变量的比例变化或百分比变化。涉及对数的回归可被用来估计比例性变化和弹性。
5. 两个变量之积被称为交叉项。当交叉项作为回归因子被包含进来时，它们允许一个变量的回归斜率依赖于另一个变量的值。

重要术语

二次回归模型　非线性回归函数　多项式回归模型　三次回归模型　弹性　指数函数　自然对数　线性对数模型　对数线性模型　双对数模型　交互作用项　交互作用回归因子　交互作用回归模型



化，再计算一下住房带有景色时的价格期望变化。两者是否存在很大的差异？这个差异在统计上是显著的吗？

表 6—4　　　　**一个特定地区在过去一年中售出的住房的数据**

因变量：ln(*Price*)					
回归因子	(1)	(2)	(3)	(4)	(5)
Size	0.00042				
	(0.000038)				
ln(*Size*)		0.69	0.68	0.57	0.69
		(0.054)	(0.087)	(2.03)	(0.055)
ln(*Size*)^2				0.0078	
				(0.14)	
Bedrooms			(0.0036)		
			(0.037)		
Pool	0.082	0.071	0.071	0.071	0.071
	(0.032)	(0.034)	(0.034)	(0.036)	(0.035)
View	0.037	0.027	0.026	0.027	0.027
	(0.029)	(0.028)	(0.026)	(0.029)	(0.030)
Pool × *view*					0.0022
					(0.10)
Condition	0.13	0.12	0.12	0.12	0.12
	(0.045)	(0.035)	(0.035)	(0.036)	(0.035)
Intercept	10.97	6.60	6.63	7.02	6.60
	(0.069)	(0.39)	(0.53)	(7.50)	(0.40)
总括性统计量					
SER	0.102	0.098	0.099	0.099	0.099
$\bar{R}^2$	0.72	0.74	0.73	0.73	0.73

注：变量定义：*Price* = 销售价格（美元）；*Size* = 房屋大小（平方英尺）；*Bedrooms* = 卧室的数量；*Pool* = 二元变量（若房子有一个游泳池为 1，若没有为 0）；*View* = 二元变量（若房子有好景色为 1，若没有为 0）；*Condition* = 二元变量（若房地产经纪商报告房屋状况良好为 1，若没有为 0）。

6.3　在阅读完本章考试成绩和班级规模的分析后，一个教育工作者评论道："以我的经验，学生的成绩取决于班级规模，但不是以你们的回归所解释的方式表达的。确切地说，当班级规模少于 20 个学生时，学生成绩好，而在班级规模超过 25 个学生时，学生成绩就会很糟糕。将班级规模降到 20 个学生以下不会有什么好处，这种关系在 20 ~ 25 个学生之间的中间区域是不变的，而当它已超过 25 时，扩大班级规模没有什么损失。"这位教育工作者描述了一个"门槛效应"，其中对于小于 20 个学生的班级规模而言，成绩是不变的；对于 20 ~ 25 个学生之间的班级规模而言，成绩发生了跳跃并保持不变；对于超过 25 个学生的班级规模而言，成绩又发生了一次跳跃。为了建立这些"门槛效应"的模型，定义二元变量：

如果 $STR<20$,那么 $STRsmall=1$;否则,$STRsmall=0$。

如果 $20\leqslant STR\leqslant 25$,那么 $STRmoderate=1$;否则 $STRmoderate=0$。

如果 $STR>25$,那么 $STRlarge=1$;否则,$STRlarge=0$。

a. 考虑回归 $TestScore_i=\beta_0+\beta_1 STRsmall_i+\beta_2 STRlarge_i+u_i$。请画出回归系数假设值与该教育工作者的评论相一致的关于 $TestScore$ 和 STR 之间关系的回归函数图形。

b. 一名研究人员试图估计出下面的回归函数:$TestScore_i=\beta_0+\beta_1 STRsmall_i+\beta_2 STRmoderate_i+\beta_3 STRlarge_i+u_i$,却发现她的计算机死机了,为什么?

*6.4 请解释你将如何用5.8节中的"方法2"计算方程(6.8)下面所讨论的置信区间。(提示:这需要用回归因子和因变量的不同定义来估计一个新的回归,见练习(5.8))

6.5 考虑回归模型:$Y_i=\beta_0+\beta_1 X_{1i}+\beta_2 X_{2i}+\beta_3(X_{1i}\times X_{2i})+u_i$。使用重要概念6.1,证明:

a. $\frac{\Delta Y}{\Delta X_1}=\beta_1+\beta_3 X_2$(在保持 X_2 不变的情况下,X_1 变化的效应)

b. $\frac{\Delta Y}{\Delta X_2}=\beta_2+\beta_3 X_1$(在保持 X_1 不变的情况下,X_2 变化的效应)

c. 如果 X_1 变化 ΔX_1 而 X_2 变化 ΔX_2,那么:

$$\Delta Y=(\beta_1+\beta_3 X_2)\Delta X_1+(\beta_2+\beta_3 X_1)\Delta X_2+\beta_3\Delta X_1\Delta X_2$$

第7章 基于多元回归的评估研究

前面的三章解释了如何用多元回归来分析数据集中变量之间的关系。在本章,我们回过头来问一下,是什么因素决定多元回归研究是可靠的还是不可靠的?我们将集中在以估计某个自变量(如班级规模)的变化对因变量(如考试成绩)的因果效应分析为目的的统计研究上。对于这样的研究,多元回归在什么时候能提供一个关于因果效应的有用估计?同样重要的是,在什么时候它不能做到这一点?

为了回答这个问题,本章提出了评估统计研究的一个一般性框架,不管它们是否使用了回归分析。这个框架依赖于内部和外部有效性的概念。如果一个研究中关于因果效应的统计推断,对所研究的总体和环境设定是有效的,那么它就是内部有效的;如果它的推断能够被推广到其他的总体和环境设定中,那么它就是外部有效的。在7.1节和7.2节,我们讨论了内部有效性和外部有效性,列出了各种可能的对内部有效性和外部有效性产生威胁的因素,并讨论了在实际中如何识别这些威胁。有些威胁还不能用迄今为止所介绍的经济计量学工具来处理,本章为本书余下章节中所研究的处理这些威胁的方法提供了一个预览。

作为阐述内部和外部有效性框架的一个例子,在7.3节我们对第4章~第6章中所提出的削减学生—教师比对考试成绩影响这项研究的内部有效性和外部有效性进行了评估。

7.1 内部有效性和外部有效性

在重要概念7.1中定义的内部有效性和外部有效性的概念,提供了一个用来评估一项统计研究或经济计量研究对回答某一感兴趣的具体问题是否有用的框架。

内部有效性和外部有效性,将所研究的总体与环境设定同结果被推广到的总体和环境设定区别开来。所谓**研究总体**(population studied),是指我们所研究的样本——个人、公司、学区等等——都是从其中抽取出来的实体性总体。所谓结果被推广到的总体或**感兴趣的总体**(population of interest),是指所研究的因果关系推断被应用于其中的实体性总体。

例如,一位高中(9~12年级)校长可能想要将我们关于加利福尼亚州小学学区(研究的总体)的班级规模和考试成绩的研究结论推广到高中的总体(感兴趣的总体)中。

对于"环境设定",我们意指制度的、法律的、社会的和经济的环境。例如,在实验室中进行的评估种植有机西红柿的方法,是否可以被推广到田地里,即在实验室的环境设定下起作用的有机方法在现实世界的环境设定下是否仍起作用,知道这一结论是非常重要的。在本节后面的部分,我们提供了总体和环境设定差异的其他例子。

7.1.1 对内部有效性的威胁

内部有效性有两个组成部分。首先,因果效应的估计量应该是无偏的且一致的。例如,如果$\hat{\beta}_{STR}$是某个回归中学生—教师比的单位变化对考试成绩影响的OLS估计量,那么$\hat{\beta}_{STR}$就应该是学生—教师比发生变化对真实总体因果效应即β_{STR}的一个无偏的且一致的估计量。其次,假设检验应该具有希望达到的显著性水平(在零假设下检验的实际拒绝率应该等于它希望达到的显著性水平),而且置信区间应该具有希望达到的置信水平。例如,如果一个置信区间被构造为$\hat{\beta}_{STR} \pm 1.96SE(\hat{\beta}_{STR})$,那么这个置信区间应该在重复抽取的样本中以95%的概率包含真实的总体因果效应β_{STR}。

重要概念 7.1

内部有效性和外部有效性

如果一项关于因果效应的统计推断对所研究的总体来说是有效的,那么该统计分析就是**内部有效的(internally valid)**。如果其推断和结论能够从所研究的总体和环境设定中,推广到其他的总体和环境设定中,那么这个分析就是**外部有效的(externally valid)**。

在回归分析中,因果效应使用所估计的回归函数进行估计,而假设检验则使用所估计的回归系数及其标准误来进行。因此,在以OLS回归为基础的一项研究中,对内部有效性的要求就是其OLS估计量是无偏的且一致的,而且标准误是在满足置信区间具有希望达到的置信水平的条件下进行计算的。许多原因使得这一点可能不会发生,而这些原因就构成了对内部有效性的威胁。这些威胁导致了重要概念5.4中一个或多个最小二乘假设的失效。例如,我们已详细讨论过的一个威胁是遗漏变量偏差,这导致了一个或多个回归因子和误差项相关,从而违背了第一条最小二乘假设条件。如果可以获得该遗漏变量的数据,那么通过将这个变量作为额外的回归因子引入到回归方程中,这个威胁就可以避免。

7.2节详细讨论了多元回归分析中内部有效性的不同威胁,以及如何缓解它们的影响。

7.1.2 对外部有效性的威胁

外部有效性的潜在威胁是由所研究的总体和环境设定与感兴趣的总体和环境设定之间的差异所引起的。

总体间的差异。所研究的总体和感兴趣的总体之间的差异可能对外部有效性构成威胁。例如,化学药品毒性效应的实验室研究典型地使用像老鼠这样的动物总体(所研究的总体),但是研究结果却被用来书写人类总体(感兴趣的总体)的健康和安全规则。老鼠和人类之间是否存在相当大的不同,以至于影响到该项研究的外部有效性,这是个有争论的问题。



更一般地说,真实的因果效应在所研究的总体和感兴趣的总体中可能是不一样的,这可

能是因为总体被选择的方式在研究总体和感兴趣总体之间是不同的，可能是因为总体的特征不同，可能是因为地理的差异，也可能是因为研究已经过时了。

环境设定间的差异。即使所研究的总体和感兴趣的总体是完全相同的，但如果环境设定(settings)不同，也不应该去推广研究结果。例如，如果两所大学规定的法定饮酒年龄不同的话，那么禁酒广告运动对大学狂欢饮酒影响的研究可能无法推广到另一个相同的大学学生团体中。在这种情形下，实施研究的总体的法律环境设定不同于其结果应用的总体的法律环境设定。



更一般地说，环境设定不同的例子包括机构环境差异(公立大学与宗教大学)、法律差异(法定饮酒年龄上的差异)或自然环境差异(南加利福尼亚州与费尔班克斯州、阿拉斯加州的狂欢饮酒会)。

在考试成绩和学生一教师比案例中的应用。第5章和第6章报告了削减学生一教师比可以改进学生考试成绩这个结果，这个结果在统计上是显著的，但实质上量却是很小的。这个分析是根据加利福尼亚州学区的数据做出的。假设这些结果在目前是内部有效的，那么，这个结果可能被推广到哪些感兴趣的其他总体和环境设定中呢?

所研究的总体和环境设定与感兴趣的总体和环境设定越接近，外部有效性就越强。例如，大学生和大学教育非常不同于小学生和初等教育，所以要把从加利福尼亚州小学学区数据中所估计的削减班级规模的影响推广到大学中，似乎是不合理的。另一方面，全美国的小学学生、课程和小学中的组织广泛地相似，因此，加利福尼亚州的研究结果可以被推广到美国其他小学学区的标准化考试成绩分析中，这似乎是合理的。

如何评价一项研究的外部有效性? 外部有效性必须使用关于所研究的总体和环境设定与感兴趣的总体和环境设定的具体知识进行判断。这两者之间的重要差异将会对该项研究的外部有效性提出质疑。

有时存在两个或多个关于不同的但却是相关的总体的研究。如果是这样的话，那么两项研究的外部有效性可以通过比较它们的结果进行检验。例如，在7.3节我们分析了马萨诸塞州小学学区的考试成绩和班级规模数据，并比较了马萨诸塞州和加利福尼亚州的结果。一般地说，两个或多个研究中的相似结论支持了外部有效性的断言，而它们的结论中不易解释的差异对它们的外部有效性提出了质疑。①

如何设计一项具有外部有效性的研究? 由于外部有效性的威胁来自于总体和环境设定缺乏可比性，因此最好在研究的早期阶段，在搜集数据之前将这些威胁最小化。研究设计超出了本书的范围，有兴趣的读者可以参考 Shadish，Cook 与 Campbell(2002)的书。

7.2 对多元回归分析内部有效性的威胁

如果所估计的回归系数是无偏的且一致的，并且它们的标准误会生成与希望达到的置信水平一致的置信区间，那么基于回归分析的研究就是内部有效的。本节将研究为什么多元回归系数的 OLS 估计量可能是有偏的五个原因(即使在大样本条件下)：遗漏变量，回归函数函数形式的误定，自变量的不精确测度("变量误差")，样本选择，联立因果关系。之所

① 对同一个主题进行的许多相关研究的比较，我们称其为元分析。例如，第5章的一般兴趣框中关于"莫扎特效应"的讨论就是以元分析为基础的。对多项研究进行元分析有其自身的挑战。你如何从差的研究中挑选出好的研究? 当因变量不同时如何对研究进行比较? 对于一项大型的研究和一项小型的研究，你会更重视大型研究的结论吗? 对元分析及其挑战的讨论超出了本书的范围，有兴趣的读者可参考 Hedges 与 Olkin(1985)以及 Cooper 与 Hedges(1994)的文章。

以会出现所有这五个偏差来源，是因为总体回归中的回归因子和误差项相关，违背了重要概念5.4中的第一个最小二乘假设条件。我们讨论了如何处理每一个偏差的来源并如何减少这些来源所造成的偏差。本节最后讨论了导致不一致标准误的情况以及如何处理这些问题的方法。

7.2.1 遗漏变量偏差

回想一下，当一个既决定 Y 又和方程中的一个或多个回归因子相关的变量从回归中遗漏掉时，就会产生遗漏变量偏差。这个偏差即使在大样本中也存在，因此在这种情况下 OLS 估计量是不一致的。如何能更好地使遗漏变量偏差最小化，取决于是否可获得潜在的遗漏变量的数据。

当遗漏变量可被观测时，解决遗漏变量偏差的办法。如果你有关于遗漏变量的数据，那么你就可以把这个变量引入到多元回归中，这样问题就解决了。然而，增加新的变量有收益也有损失。一方面，遗漏了变量可能会导致遗漏变量偏差。另一方面，当它不属于回归时（也就是说，当它的总体回归系数为零时），引人这个变量会降低其他回归系数估计量的精度。换句话说，决定是否引入一个变量，涉及在发生偏差和增加系数方差之间进行权衡。在实践中，有四个步骤可以帮助你决定是否在回归中引入一个变量或一组变量。

第一步，识别回归中那个系数是关键系数。在考试成绩的回归中，关键系数就是学生—教师比的系数，因为最初提出的问题与减少学生—教师比对考试成绩的影响有关。

第二步，问问你自己：在这个回归中，重要的遗漏变量偏差的最有可能的来源是什么？回答这个问题要求应用经济理论和专家知识，而且在你实际进行任何回归以前就应该想到。因为这在分析数据之前就应该做了，所以又被称为先验（"事前的"）推理。在考试成绩的例子中，这一步需要我们识别那些决定考试成绩的因素，如果被忽略了，就可能会使班级规模效应的估计量产生偏差。这一步的结果引出了基准的回归设定，也是经验回归分析的起点，可能会有助于缓和遗漏变量偏差的一系列额外的"可疑"变量。

第三步，用第二步中识别出来的额外可疑变量来扩展你的基准设定，并检验它们的系数均为零的假设。如果额外变量的系数在统计上是显著的，或者如果当额外变量被包含进来时，感兴趣的估计系数发生了明显变化，那么它们应该保留在设定中，你应该修正基准设定。如果不是的话，这些变量就可以排除在回归以外。

第四步，以表格的形式把你的分析结果准确地概括出来。这向潜在的抱怀疑态度的人提供了一种"完全的披露"，于是他（或她）可以据此得出自己的结论。表5—2和表6—3就是这个方法的例子。例如，在表6—3中，我们本可以只介绍第(7)列中的回归，因为该回归概括了该表中其他回归的相关效应和非线性问题，可是，给出其他的回归结果，就可以允许那些持怀疑意见的读者得出自己的结论。

这些步骤在重要概念7.2中总结。

当遗漏变量观测不到时，解决遗漏变量偏差的方法。如果你没有关于遗漏变量的数据，那么将该遗漏变量添加到回归中就不是一种很好的选择，不过仍然有其他三种解决遗漏变量偏差的方法。这三种解决方法中的每个方法，通过使用不同类型的数据，都可以防止遗漏变量偏差的发生。

第一种解决方法是，使用相同的观测单位在不同时点上被适时观测到的数据。例如，对同一个地区而言，可以搜集1995年的考试成绩和相关的数据，同时在2000年再搜集一次，这种形式的数据叫做面板数据。如第8章所解释的，只要那些遗漏变量不随时间变化而变

化，面板数据就会使控制那些无法观测到的遗漏变量成为可能。

第二种解决方法是，使用工具变量回归。这种方法依赖于一个新的变量，这个变量被称为工具变量。工具变量回归在第10章中进行讨论。

第三种解决方法是，使用一项研究设计，在这项研究设计中，使用随机化的控制实验来研究我们所关心的效应（比如，减少班级规模对学生成绩的影响）。随机化控制实验在第11章中讨论。

7.2.2 回归函数中函数形式的误设定

如果真实的总体回归函数是非线性的，但所估计的回归却是线性的，那么这个函数形式误定（functional form misspecification）就会使OLS估计量变成是有偏的。这种偏差是遗漏变量偏差的一种类型，其中的遗漏变量是反映回归函数中缺失的非线性特征的项。例如，如果总体回归函数是个二次多项式，那么如果遗漏了自变量平方项，相应的回归就将产生遗漏变量偏差。

函数形式误定的解决方法。当自变量（如考试成绩）为连续的时，那么可以用第6章的方法解决这个潜在的非线性问题。不过，如果自变量是离散的或二元的（例如，如果第i个人上了大学，那么Y_i等于1；否则等于0），那么问题就更复杂了。第9章讨论了离散型因变量的回归。

重要概念7.2

我应该在回归中包含更多的变量吗

如果你在多元回归中包含了另一个变量，那么你会消除因排除那个变量所引起的遗漏变量偏差的可能性，但是你所关心的系数估计量的方差可能会增加。这里给出一些帮助你决定是否包含额外变量的准则：

1. 明确确定系数或你所感兴趣的系数。

2. 使用先验推理，识别遗漏变量偏差的最重要的潜在来源，导出基准设定和一些“可疑”变量。

3. 检验额外的可疑变量的系数是否为零。

4. 用“完全披露”的有代表性的列表方法给出你的分析结果，以便其他人能够看到引入可疑变量对所关心的系数的效应。如果你引入了一个可疑变量，你的结果发生变化了吗？

7.2.3 变量误差

假设在考试成绩对学生—教师比的回归中，我们不经意地将数据混合在一起了，使我们以五年级的考试成绩对该地区十年级的学生—教师比进行了回归。尽管五年级的学生—教师比和十年级的学生—教师比可能是相关的，但它们是不一样的，因此这种混合会导致估计系数的偏差，这就是变量误差偏差（errors-in-variables bias）的一个例子，因为它的来源是自变量测度误差。这个偏差即使在很大的样本中依然存在，因此如果存在测度误差的话，那么OLS估计量就是不一致的。

测度误差有许多可能的来源。如果数据是通过调查搜集的，那么应答者可能会给出错误的回答。例如，在当前人口调查中有一个问题涉及去年的收人。应答者可能不知道他的确切收人，或者可能由于某些其他的原因说错了。如果数据是从计算机化的管理记录中获得的，那么当数据被初次输入时，可能也会有打字输人错误。

为了理解变量误差偏差导致回归因子和误差项之间的相关，假设有一个回归因子 X_i（如实际收入），但是 X_i 是由 $\widetilde{X}_i$（应答者的收入的估计值）不精确地测度出的。因为观测到的是 $\widetilde{X}_i$ 而非 X_i，所以实际上所估计出来的回归方程是基于 $\widetilde{X}_i$ 的回归。用这个不精确测度到的变量 $\widetilde{X}_i$，写出总体回归方程 $Y_i = \beta_0 + \beta_1 X_i + u_i$：

$$\begin{aligned} Y_i &= \beta_0 + \beta_1 \widetilde{X}_i + [\beta_1 (X_i - \widetilde{X}_i) + u_i] \\ &= \beta_0 + \beta_1 \widetilde{X}_i + v_i \end{aligned} \tag{7.1}$$

其中，$v_i = \beta_1 (X_i - \widetilde{X}_i) + u_i$，因此，以 $\widetilde{X}_i$ 的形式表示的总体回归方程有个包含 X_i 和 $\widetilde{X}_i$ 之差的误差项。如果这个差与测度值 $\widetilde{X}_i$ 相关，那么回归因子 $\widetilde{X}_i$ 将会与误差项相关，$\hat{\beta}_1$ 将会是有偏的且不一致的。

$\hat{\beta}_1$ 偏差确切的大小和方向依赖于 X_i 和 $(X_i - \widetilde{X}_i)$ 之间的相关性，而这个相关性反过来又依赖于测度误差的具体性质。

举一个例子，假设调查应答者给出了自变量 X_i 的实际值的最好猜测或回忆，数学上描述这一点的简便方法是假设 X_i 的测度值等于实际的非测度值加上一个完全随机成分 w_i，因此，该变量的测度值（记为 $\widetilde{X}_i$）为 $\widetilde{X}_i = X_i + w_i$。由于误差项是完全随机的，因此我们可以假定 w_i 具有零均值和方差 σ_w^2，并且 w_i 同 X_i 和回归误差项 u_i 不相关。在这个假设下，应用一点儿代数知识[①]就可以证明 $\hat{\beta}_1$ 具有概率极限：

$$\hat{\beta}_1 \xrightarrow{p} \frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2} \beta_1 \tag{7.2}$$

也就是说，如果这种测度不精确性具有只对自变量的实际值增加一个随机成分的作用，那么 $\hat{\beta}_1$ 就是不一致的。因为比率 $\frac{\sigma_X^2}{\sigma_X^2 + \sigma_w^2}$ 小于 1，所以 $\hat{\beta}_1$ 将会向 0 方向发生偏差，即使在大样本中亦如此。在这种极端情形下，即测度误差很大，以至于基本上没有留下关于 X_i 的任何信息，在表达式(7.2)最后表达式中的方差比率为零，$\hat{\beta}_1$ 依概率收敛于零。在另一个极端情形下，当不存在测度误差时，$\sigma_w^2 = 0$，所以 $\hat{\beta}_1 \xrightarrow{p} \beta_1$。

尽管表达式(7.2)中的结果对这个特殊类型的测度误差来说是具体的，但它阐明了更一般的命题——如果自变量没有被精确地测度，那么即使在大样本中 OLS 估计量也是有偏的。变量误差偏差在重要概念 7.3 中总结。

变量误差偏差的解决方法。解决变量误差问题的最好方法是得到 X 的精确测度值。如果这是不可能的，那么还有一些经济计量方法可用来缓和变量误差的偏差。

第一种方法是工具变量回归。这依赖于另一个与实际值 X_i 相关，但与测量误差不相关的变量（“工具”变量）。这种方法在第 10 章中研究。

第二种方法是导出测度误差的数学模型，而且如果可能的话，用相应导出的公式来调整



---

① 在这个测度误差假设下，$v_i = \beta_1 (X_i - \widetilde{X}_i) + u_i = -\beta_1 w_i + u_i$，$\mathrm{cov}(\widetilde{X}_i, u_i) = 0$，且 $\mathrm{cov}(\widetilde{X}_i, w_i) = \mathrm{cov}(X_i + w_i, w_i) = \sigma_w^2$，所以 $\mathrm{cov}(\widetilde{X}_i + v_i) = -\beta_1 \mathrm{cov}(\widetilde{X}_i + w_i) + \mathrm{cov}(\widetilde{X}_i + u_i) = -\beta_1 \sigma_w^2$。因此，由表达式(5.1)得，$\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / \sigma_{\widetilde{X}}^2$。现有 $\sigma_{\widetilde{X}}^2 = \sigma_X^2 + \sigma_w^2$，所以 $\hat{\beta}_1 \xrightarrow{p} \beta_1 - \beta_1 \sigma_w^2 / (\sigma_X^2 + \sigma_w^2) = [\sigma_X^2 / (\sigma_X^2 + \sigma_w^2)] \beta_1$。

估计值。例如，如果一位研究人员相信，所测度的变量实际上是实际值和随机测度误差项之和，而且如果她知道或能够估计出比率 σ_w^2/σ_X^2，那么她就能够用表达式(7.2)计算一个修正了向下偏差的 β_1 的估计量。由于这种方法要求知道关于测度误差性质的专业知识，对于给定的一个数据集及其测度问题，细节通常是很具体的，因此我们在本书中将不再继续研究这种方法。

重要概念 7.3

变量误差偏差

当一个自变量被不精确地测度时，就会产生 OLS 估计量的变量误差的偏差。这个偏差依赖于测度误差的性质，并且即使在样本容量很大的情况下依然存在。如果所测度的变量等于实际值加上一个零均值的、独立分布的测度误差项，那么在只具有一个右边变量的回归中，OLS 估计量偏向于零，表达式(7.2)给出了它的概率极限。

7.2.4 样本选择

当数据的可获得性受到选择过程的影响，而且这个选择过程与自变量的值相关时，就会发生样本选择偏差(sample selection bias)。这个选择过程会引致误差项和回归因子之间的相关性，从而会导致 OLS 估计量的偏差。

与因变量的值不相关的样本选择不会产生偏差。例如，如果数据是通过简单随机抽样从总体中搜集的，那么抽样方法(从总体中随机地抽取)与自变量的值没有关系。这样的抽样不会引人偏差。

当抽样方法与因变量的值相关时，就可能会引人偏差。在第 3 章的方框中给出了一个关于民意测验中样本选择偏差的例子。在那个例于中，样本选择方法(随机选择汽车车主的电话号码)与因变量(个人支持谁在 1936 年当总统)相关，因为在 1936 年拥有电话的汽车车主比较可能是共和党人。

在经济学中，样本选择偏差的一个例于，经常是用工资对教育的回归中估计每增加一年的教育对工资的影响作为说明。根据定义，只有那些有工作的人才有工资。决定一个人是否有工作的因素(可观测的和不可观测的)——教育、经验、居住地、能力、运气等——与决定该人在有工作时赚多少钱的因素相似。因此，某人有工作的事实表明，如果其他所有的条件都相同，那么他的工资方程中的误差项是正的。换句话说，某人是否有工作部分地由工资回归方程误差项中的遗漏变量决定。因此，某人有工作，而且显示在数据集中这一简单事实，至少平均来看，提供了误差项为正的，而且可能与回归因子相关的信息。这也会使 OLS 估计量产生偏差。

样本选择偏差在重要概念 7.4 中总结。

选择偏差的解决方法。到目前为止，我们所讨论的方法还不能消除样本选择偏差。估计带有样本选择的模型的方法超出了本书的范围，这些方法建立在第 9 章所介绍的技术之上，那一部分提供了进一步的参考。

7.2.5 联立因果关系

迄今为止，我们是假设因果关系从回归因子传导到因变量(即 X 引致 Y)，但是如果因果关系也是从因变量传导到一个或多个回归因子(即 Y 引致 X)，那么情况会怎样呢？如果是这样的话，因果关系除了“向后”传导外，还“向前”传导，也就是说，存在联立因果关系

(simultaneous causality)。如果存在联立因果关系,那么 OLS 回归得到两个效应,因此 OLS 估计量是有偏的且不一致的。

重要概念 7.4

样本选择偏差

当选择过程影响到数据的可获得性且该过程与因变量相关时,就会产生样本选择偏差。样本选择引致一个或多个回归因子与误差项相关,从而导致 OLS 估计量的偏差且不一致性。

例如,我们对考试成绩的研究集中于分析减少学生一教师比对考试成绩的影响,所以因果关系被假定为是从学生一教师比传导到考试成绩。然而,假设一个政府最初资助在考试成绩差的学区雇佣教师,如果是这样的话,因果关系将是双向的:按通常的教育推理,低的学生一教师比会导致高的考试成绩,但是由于该项政府资助计划,低的考试成绩也会导致低的学生一教师比。

联立因果关系导致了回归因子与误差项相关。在考试成绩的例子中,假设有一个遗漏因素导致了比较差的考试成绩,由于该项政府计划,这个产生低考试成绩的因素反过来又导致低的学生一教师比。因此,在考试成绩对学生一教师比的总体回归中,负的误差项降低了考试成绩,但是由于该项政府计划的作用,它还导致了学生一教师比的降低。换句话说,在总体回归中学生一教师比与误差项是正相关的,这反过来又导致了联立因果关系偏差和 OLS 估计量的不一致性。

在数学上,引入另外一个描述反向因果关系的方程,能够使误差项与回归因子之间的这个相关关系变得精确。为简便起见,只考虑两个变量 X 和 Y,而忽略其他可能的回归因子,因此只有两个方程,一个方程是 X 引致 Y,而另一个方程是 Y 引致 X:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{7.3}$$

$$X_i = \gamma_0 + \gamma_1 Y_i + v_i \tag{7.4}$$

公式(7.3)是我们所熟悉的方程,其中 β_1 是 X 的变化对 Y 的影响,这里 u 代表其他因素。公式(7.4)表示 Y 对 X 的反向的因果效应。在考试成绩这个问题中,公式(7.3)表示班级规模对考试成绩的教育效应,而公式(7.4)表示由政府计划所引致的考试成绩对班级规模的反向因果效应。

联立因果关系导致了公式(7.3)中 X_i 与误差项 u_i 的相关。为了弄明白这一点,假设 u_i 是负的,它使 Y_i 减少。但是,这个较低的 Y_i 值会通过第二个方程影响 X_i 的值,而且如果 γ_1 是正的,那么低的 Y_i 值会导致一个低的 X_i 值。因此,如果 γ_1 是正的,那么 X_i 和 u_i 将会是正相关的。[①]

因为这在数学上可以用两个联立方程系统来表达,所以联立因果关系偏差有时被称为**联立方程偏差(simultaneous equations bias)**。联立因果关系偏差在重要概念 7.5 中总结。

联立因果关系偏差的解决方法。有两种缓和联立因果关系偏差的方法:一种是使用工具变量回归,也即第 10 章的主题;另一种是设计并实施一项随机化控制实验,在该试验中反向因果关系渠道被设为无效,这样的实验在第 11 章中讨论。

---

① 为了在数学上证明这一点,注意公式(7.4)隐含着 $\text{cov}(X_i, u_i) = \text{cov}(\gamma_0 + \gamma_1 Y_i + v_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) + \text{cov}(v_i, u_i)$。假设 $\text{cov}(v_i, u_i) = 0$,根据公式(7.3),这又意味着 $\text{cov}(X_i, u_i) = \gamma_1 \text{cov}(Y_i, u_i) = \gamma_1 \text{cov}(\beta_0 + \beta_1 X_i + u_i, u_i) = \gamma_1 \beta_1 \text{cov}(X_i, u_i) + \gamma_1 \sigma_u^2$。解 $\text{cov}(X_i, u_i)$ 得到 $\text{cov}(X_i, u_i) = \gamma_1 \sigma_u^2 / (1 - \gamma_1 \beta_1)$。

7.2.6 OLS 标准误不一致性的来源

不一致的标准误对内部有效性提出了不同的威胁。即使 OLS 估计量是一致的,样本很大,不一致的标准误也将会导致规模不同于希望得到的显著性水平的假设检验,导致在95%的重复样本中不会包含真实值的“95%”的置信区间。

存在不一致的标准误主要有两个原因:对异方差的不恰当的处理和观测值之间误差项的相关性。

异方差。在4.9节中我们已讨论过,由于历史的原因,一些回归软件只报告同方差惟一的标准误。然而,如果回归误差是异方差的,那么,那些标准误就不是假设检验和置信区间估计的可靠基础。这个问题的解决办法是,使用异方差稳健性标准误并使用异方差稳健性的方差估计量来构造 F 统计量。在现代的软件包中,异方差稳健性标准误是作为一个选项给出的。



重要概念 7.5

联立因果关系偏差

联立因果关系偏差,又叫联立方程偏差。在 Y 对 X 的回归中,在除了我们所关心的从 X 到 Y 的因果关系以外,还存在从 Y 到 X 的因果关系时,就会出现联立因果关系偏差。这种反向的因果关系,使得 X 与我们所研究的总体回归中的误差项相关。

不同观测值之间误差项的相关性。在某些设定中,总体回归误差在不同的观测值之间可能是相关的。如果数据是从总体中随机抽取的,那么这就不会发生,因为抽样过程的随机性确保了误差项在不同的观测值之间是独立分布的。不过,抽样有时候只是部分随机的。最一般的情况是,对相同实体的数据随时间变化重复地观测,例如,在不同年份对同一学区进行重复观测。如果构成回归误差的遗漏变量具有持久性影响(如地区人口统计特征),那么这会引起随时间变化的回归误差的“序列”相关。另一个例子是当抽样以地理单位为基础时。如果存在反映地理影响的遗漏变量,那么这些遗漏的变量会导致相邻观测值之间回归误差的相关性。

不同观测值之间回归误差的相关性不会使 OLS 估计量产生有偏性或者是不一致性,但是它确实违背了重要概念 5.4 中的第二条最小二乘假设条件。结果是,OLS 标准误——不管是同方差惟一的标准误还是异方差稳健的标准误——是不正确的,因为它们不能生成我们希望达到的置信水平的置信区间。

在许多情形下,这个问题可以通过使用标准误差的一个替代公式来解决。我们给出了一个计算标准误的公式,它对第 12 章时间序列数据回归中讨论的异方差和序列相关而言都是稳健的。

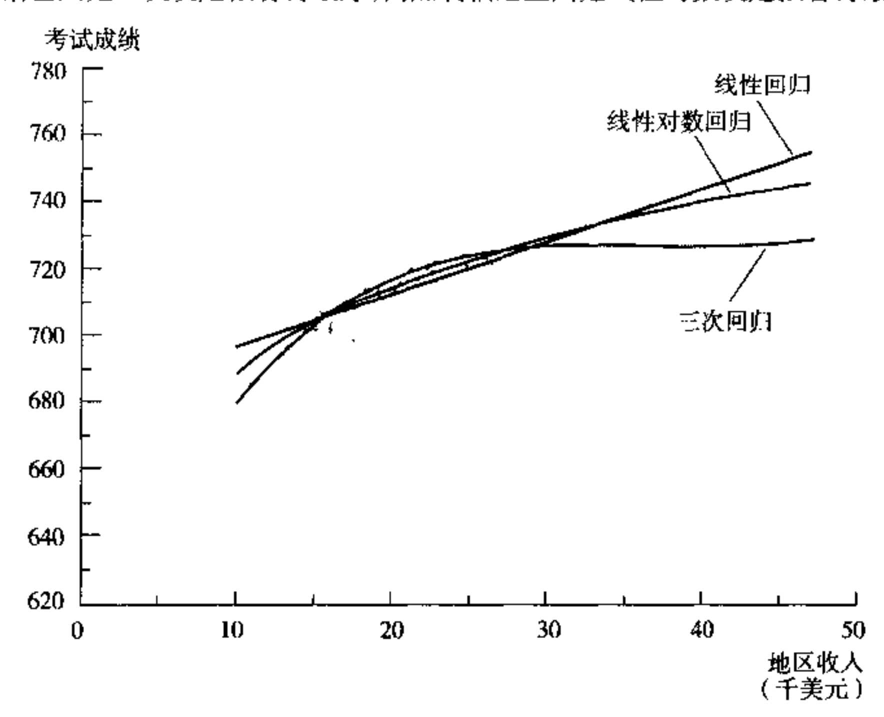
7.3 例子:考试成绩和班级规模

内部有效性和外部有效性的内容框架,可以帮助我们用批判性的眼光看一看,我们从对加利福尼亚州考试成绩数据的分析中已经学会了哪些,以及我们还有哪些没有学会。

7.3.1 外部有效性

加利福尼亚州的分析能否被推广,即它是否是外部有效的,依赖于被推广到其中的总体

一般模式也存在于马萨诸塞州的数据中。然而，描述这个非线性的最佳的函数形式却不相同，马萨诸塞州是三次设定拟合得最好，而加利福尼亚州是线性对数设定拟合得最好。



注：所估计的线性回归函数没有捕捉到马萨诸塞州数据中收入和考试成绩之间的非线性关系。对地区收入在13 000美元～30 000美元之间的包含大部分观测值的地区而言，所估计的线性对数和三次回归函数是相似的。

图7—1　马萨诸塞州数据中考试成绩与收入间的关系

多元回归结果。马萨诸塞州数据的回归结果在表7—2中给出。表的第(1)列报告了第一个回归，即只把学生—教师比作为回归因子。斜率是负的(－1.72)，并且系数为零的假设在1%的显著性水平下被拒绝($t=-1.72/0.50=-3.44$)。

表中其余的列，报告了包含控制学生特征的额外变量和将非线性引入到所估计的回归函数中的分析结果。控制英语学习者的百分比、享受免费午餐计划学生的百分比之后，地区平均收入使学生—教师比的估计系数降低了60%，从回归(1)中的－1.72到回归(2)中的－0.69和回归(3)中的－0.64。

比较回归(2)和回归(3)中的$\bar{R}^2$值，说明即使在保持学生—教师比不变的情况下，三次设定(3)提供了一个比对数设定(2)更好的关于考试成绩和收入之间关系的模型。在统计上，不存在考试成绩和学生—老师比之间非线性关系的显著证据：回归(4)中检验STR^2和STR^3的总体系数是否都为零的F统计量的p值为0.641。同理，在统计上也并不存在充分的证据表明降低学生—老师比在英语学习者很多的地区和很少的地区之间具有不同的影响(回归(5)中$HiEL \times STR$的t统计量为$0.80/0.56=1.43$)。最后，回归(6)表明，当英语学习者的百分比(它在回归(3)中是不显著的)被剔除掉时，学生—教师比的估计系数没有发生实质性的变化。简而言之，在表7—2中，回归(3)的结果对回归(4)～回归(6)中所考虑的函数形式和设定的变化不敏感。因此，我们采用回归(3)作为分析马萨诸塞州数据中学生—教师比的变化对考试成绩影响的基准估计方法。

马萨诸塞州分析结果与加利福尼亚州分析结果的比较。通过对加利福尼亚州数据的研究，我们发现：

表 7—2　　学生—教师比和考试成绩的多元回归估计(数据来自于马萨诸塞州)

因变量:学区内四年级学生的英语、数学和科学的综合平均考试成绩;220 个观测值

回归因子	(1)	(2)	(3)	(4)	(5)	(6)
学生—教师比	-1.72**	-0.69*	-0.64*	12.4	-1.02**	-0.67*
(*STR*)	(0.50)	(0.27)	(0.27)	(14.0)	(0.37)	(0.27)
(STR^2)				-0.680		
				(0.737)		
(STR^3)				0.011		
				(0.013)		
英语学习者的百分比		-0.411	-0.437	-0.434		
		(0.306)	(0.303)	(0.300)		
英语学习者的百分比 > 中位数?					-12.6	
(二元变量,*HiEL*)					(9.8)	
HiEL × *STR*					0.80	
					(0.56)	
享受免费午餐计划		-0.521**	-0.582**	-0.587**	-0.709**	-0.653**
的学生的百分比		(0.077)	(0.097)	(0.104)	(0.091)	(0.72)
地区收入(对数)		16.53**				
		(3.15)				
地区收入			-3.07	-3.38	-3.87*	-3.22
			(2.35)	(2.49)	(2.49)	(2.31)
地区收入^2			0.164	0.174	0.184*	0.165
			(0.085)	(0.089)	(0.090)	(0.085)
地区收入^3			-0.0022*	-0.0023*	-0.0023*	-0.0022*
			(0.0010)	(0.0010)	(0.0010)	(0.0010)
截距	739.6**	682.4**	744.0**	665.5**	759.9**	747.4**
	(8.6)	(11.5)	(21.3)	(81.3)	(23.2)	(20.3)
检验剔除变量组 *F* 统计量和 *p* 值						
所有 *STR* 变量				2.86	4.01	
和交叉项 = 0				(0.038)	(0.020)	
STR^2, STR^3 = 0				0.45		
				(0.641)		
$Income^2$, $Income^3$ = 0			7.74	7.75	5.85	6.55
			(< 0.001)	(< 0.001)	(0.003)	(0.002)
HiEL, *HiEL* × *STR*					1.58	
					(0.208)	
SER	14.64	8.69	8.61	8.63	8.62	8.64
$\bar{R}^2$	0.063	0.670	0.676	0.675	0.675	0.674

注:这些回归是用附录 7.1 中所描述的马萨诸塞州小学学区的数据进行估计的。标准误在系数下面的括号中给出,*p* 值在 *F* 统计量下面的括号中给出。单个系数在*5% 或**1% 的水平下在统计上是显著的。

a. 添加控制学生背景特征的变量使学生—教师比的系数从 -2.28(表5—2 中的回归(1))降到 -0.73(表6—3 中的回归(2)),降低了68%。

b. 学生—教师比的真实系数为零的假设在1%的显著性水平下被拒绝,即使在添加了控制学生背景和地区经济特征的变量之后亦如此。

c. 削减学生—教师比的效应并不十分依赖于该地区英语学习者的百分比。

d. 有些证据表明考试成绩和学生—教师比之间的关系是非线性的。

在马萨诸塞州的数据中我们是否发现了相同的结论呢?对于结论(a)、(b)和(c),答案是肯定的。包含额外的控制变量使学生—教师比的系数从 -1.72(表7—2 中的回归(1))降到 -0.69(表7—2 中的回归(2)),降低了60%。学生—教师比的系数在添加了控制变量后仍然显著。在马萨诸塞州的数据中,那些系数仅在5%的水平下是显著的,而在加利福尼亚州的数据中,它们在1%的水平下是显著的。然而,加利福尼亚州的数据有近两倍于马萨诸塞州的观测值,所以加利福尼亚州的估计值更精确就不足为奇了。和在加利福尼亚州的数据中一样,在马萨诸塞州的数据中,不存在关于学生—教师比和表示地区英语学习者百分比大小的二元变量之间交互作用的在统计上的显著性证据。

可是,结论(d)在马萨诸塞州的数据中并不成立:当针对三次设定进行检验时,学生—教师比和考试成绩之间的关系是线性的这一假设在5%的显著性水平下不能被拒绝。

由于这两个标准化考试是不同的,因此系数本身不能直接进行比较:马萨诸塞州考试的1分与加利福尼亚州考试的1分是不同的。不过,如果将考试成绩转化为相同的单位,那么所估计的班级规模效应就能够进行比较了。如此做的一种方法是将考试成绩进行标准化转换,即分数减去样本平均值,然后再除以标准差,所得结果具有均值为0和方差为1的特征。在对转换后的考试成绩的回归中,斜率系数等于初始回归中的斜率系数除以考试成绩的标准差。因此,除以考试成绩标准差之后的学生—教师比的系数就可以在两个数据集之间进行比较了。

这个比较在表7—3 中进行。表的第1列报告了在具有包含英语学习者的百分比、享受免费午餐计划学生的百分比和平均地区收入作为控制变量的回归中的学生—教师比系数的 OLS 估计值。第2列报告了考试成绩在各地区间的标准差。最后两列报告了学生—教师比降低2个单位(即我们教育主管的建议)对考试成绩的估计效应,第1列用考试成绩为单位进行表示,第2列用标准差为单位进行表示。对于线性设定,使用加利福尼亚州数据的 OLS 系数估计值为 -0.73,因此以学生—教师比削减2个单位会使地区考试成绩提高 $(-0.73)\times(-2)=1.46$ 分。考试成绩的标准差是19.1分,它对应于 $1.46/19.1=0.076$ 个地区间考试成绩分布的标准差。这个估计值的标准误为 $0.26\times2/19.1=0.027$。非线性模型的估计效应及其标准误用6.1 节中所介绍的方法进行计算。

根据使用加利福尼亚州数据的线性模型,估计学生—教师比减少2个单位会使考试成绩提高0.076个标准差单位,其标准误为0.027。加利福尼亚州数据的非线性模型表明了稍微大一些的效应,其具体效应依赖于初始的学生—教师比。基于马萨诸塞州的数据,这个估计效应为0.085个标准差单位,其标准误为0.036。

这些估计值基本上是相同的,预测削减学生—教师比会提高考试成绩,但是预测的改进幅度很小。例如,在加利福尼亚州的数据中,在中位数地区和第75位百分位数地区之间的考试成绩之差为12.2分(表4—1)或0.64(12.2/19.1)个标准差。来自于线性模型的估计效应仅仅超过这个规模的1/10,换句话说,根据这个估计值,学生—教师比削减2个单位只会使该地区在地区间考试成绩分布的中位数到第75位百分位数之间的方向上移动1/10。

变量误差。地区平均学生—教师比是班级规模的一个宽泛的、不精确的潜在测度。例如，由于学生经常在地区间迁移，学生—教师比可能不准确地描述了参加考试的学生所经历的实际班级规模，它反过来会导致所估计的班级规模效应偏向于零。另一个具有潜在测量误差的变量是地区平均收入。那些数据取自于1990年的人口普查，而其他数据取自于1998年（马萨诸塞州）和1999年（加利福尼亚州）。如果该地区的经济结构在20世纪90年代发生了巨大变化，那么这将会是实际地区平均收入的不精确测度。

选择。加利福尼亚州和马萨诸塞州的数据涵盖了州内所有满足最低规模约束的公立小学学区，因此没有理由认为这里的样本选择有问题。

联立因果关系。如果标准化考试的成绩影响学生—教师比，那么联立因果关系就会出现。例如，如果某些官僚机构或政治机制对成绩差的学校或地区增加资助，这会导致这些学校和地区雇佣更多的教师，那么联立因果关系就可能发生。在这些考试期间内，马萨诸塞州不存在使学校资助均等化的机制。在加利福尼亚州，一系列法律案例表明存在一些资助均等化现象，但是这种资助的再分配并不是以学生的成绩为基础的。因此，在马萨诸塞州和加利福尼亚州，联立因果关系看上去并不是个问题。

异方差以及不同观测值之间误差项的相关性。这里和前面章节中给出的所有结果都使用了异方差稳健的标准误，因此，异方差不会威胁到内部有效性。然而，由于没有使用简单随机抽样（样本由州内所有的小学学区组成），因此观测值之间误差项相关可能会威胁到标准误的一致性。尽管存在能够应用于这种情形的其他形式的标准误公式，但是细节过于复杂且专业化，我们将它们留在更高级的课本中学习。

7.3.3 讨论和含义

马萨诸塞州和加利福尼亚州的分析结果之间的相似性表明，这些研究结论是外部有效的，也就是说，主要的结论能被推广到美国其他小学学区的标准化考试成绩中。

通过控制学生背景、家庭经济背景和地区的富裕程度，通过检查回归函数中的非线性特征，对内部有效性的最重要的一些潜在威胁已被解决了。不过，一些潜在的对内部有效性的威胁还仍然存在。一个主要的可能威胁是遗漏变量偏差，它可能会因为控制变量没有捕捉到学区的其他特征或课外学习机会而产生。

根据加利福尼亚州和马萨诸塞州的数据，我们能够解答4.1节中教育主管的问题：在控制了家庭经济背景、学生特征和地区的富裕程度之后，且在建立了回归函数的非线性模型之后，预测以2个单位的幅度削减学生—教师比，会使考试成绩大约提高0.08个地区间考试成绩分布的标准差。这个效应在统计上是显著的，但是它非常小。这个小的估计效应与许多已经完成的班级规模降低对考试成绩效应的研究结果[①]相一致。

教育主管现在能使用这个估计值来帮助她决定是否减小班级规模。在做这个决策时，她需要权衡预计减少的成本和利益。成本包括教师的工资和增加教室的费用，利益包括提高学习成绩，我们已用标准化考试成绩来测度了，但是还有其他潜在的利益我们没有研究，包括较低的辍学率和提高未来收入。所建议的缩减对标准化考试成绩的估计效应是她计算成本和利益的一个重要的考虑因素。

① 如果对了解更多的班级规模和考试成绩之间的关系感兴趣，参看Ehrenberg，Brewer，Gamoran与Willms的评论(2001a，2001b)。

7.4 结论

内部有效性和外部有效性的概念为评估从经济计量学研究中所学到的东西提供了一个框架。

如果估计的系数是无偏的、一致的，而且标准误也是一致的，那么基于多元回归研究得出的结论就是内部有效的。对这种研究的内部有效性产生威胁的因素包括遗漏变量、函数形式误定（非线性）、自变量的不精确测量（变量误差）、样本选择及联立因果关系。每一种威胁都会引起回归因子与误差项之间的相关，这反过来会使 OLS 估计量有偏且不一致。如果观测值之间误差项相关，就像它们在时间序列数据中可能会的那样，或如果它们是异方差的，但是标准误却是用同方差惟一的公式计算的，那么内部有效性就受到了损害，因为标准误将会是不一致的。后面的这些问题可通过恰当地计算标准误来解决。

如果一项研究的发现可被推广到所研究的总体和环境设定之外，那么和任何统计研究一样，使用回归分析的该项研究是外部有效的。有时，它能有助于比较两个或多个同一主题的研究。然而，不论是否存在两个或多个这样的研究，评价外部有效性需要对所研究的总体和环境设定与结论被推广到的总体和环境设定的相似性进行判断。

本书接下来的两部分将提出解决那些不能单独由多元回归分析来缓和的对内部有效性有威胁的方法。第 3 部分以缓和 OLS 估计量中所有这五个潜在偏差来源为目的扩展了多元回归模型；第 3 部分还讨论了一种获得内部有效性和随机化控制实验的不同方法。第 4 部分提出了分析时间序列数据和使用时间序列数据来估计所谓的动态因果效应（它是随时间变化的因果效应）的方法。

总结

1. 评价一项统计研究成果，可通过询问该成果是否达到内部有效和外部有效来实现。如果一项对因果效应的统计推断对所研究的总体而言是有效的，那么该项研究是内部有效的。如果该推断和结论能从所研究的总体和环境设定推广到其他的总体和环境设定，那么该项研究是外部有效的。

2. 在回归分析中，对内部有效性存在两个主要的威胁：第一，如果回归因子与误差项是相关的，那么 OLS 估计量将会是不一致的；第二，当标准误不正确时，置信区间和假设检验是无效的。

3. 当存在遗漏变量、不正确的函数形式、一个或多个回归因子有测量误差、样本不是从总体中随机选择的，或者回归因子与因变量之间存在联立因果关系时，回归因子和误差项之间就可能相关。

4. 当误差项是异方差的，而计算机软件却使用同方差惟一的标准误时，或者当误差项在不同的观测值之间相关时，标准误就是不正确的。



重要术语

研究总体　感兴趣的总体　内部有效性　外部有效性　函数形式误定　变量误差偏差　样本选择偏差　联立因果关系　联立方程偏差

复习概念

7.1 内部有效性与外部有效性之间的区别是什么？研究的总体和感兴趣的总体之间的区别又是什么？

7.2 根据偏差与方差之间的权衡，重要概念7.2解释了变量选择问题。这个权衡是什么意思？为什么包含一个额外的回归因子能够降低偏差？为什么会增加方差？

7.3 经济变量经常带有测量误差，这是否意味着回归分析是不可靠的？请说明理由。

7.4 假设一个州向州内所有的三年级学生提供自愿的标准化考试，而且这些数据被用于班级规模对学生成绩影响的研究中。请解释样本选择偏差是怎样使研究结果变得无效的。

7.5 一名研究人员利用城市水平的数据估计了警察经费支出对于犯罪率的影响。请解释联立因果关系是怎样使研究结果变得无效的。

7.6 一名研究人员使用两个不同的软件包估计同一个回归。第一个软件包使用了同方差惟一的公式计算标准误，第二个软件包使用了异方差稳健的公式计算标准误。这两个标准误非常不同，你该使用哪一个？为什么？

练习

*7.1 假设你刚阅读了一份非常细致的关于广告宣传对香烟需求效应的统计研究。该研究使用的是来自于纽约州20世纪70年代期间的数据，研究结论是：公共汽车和地铁上的广告比印刷广告更有效。使用外部有效性的概念来确定这些结果是否有可能被用于20世纪70年代的波士顿地区、20世纪70年代的洛杉矶、2002年的纽约。

7.2 考虑单变量回归模型：$Y_i=\beta_0+\beta_1X_i+u_i$，并假设它满足重要概念4.3中的假设。设$Y_i$有测量误差，因此数据是$\tilde{Y}_i=Y_i+w_i$，这里的$w_i$是测量误差，它是独立同分布的，且独立于$Y_i$和$X_i$。考虑总体回归函数$\tilde{Y}_i=\beta_0+\beta_iX_i+v_i$，这里$v_i$是使用了错误测量的因变量$\tilde{Y}_i$的回归误差。

a. 证明：$v_i=u_i+w_i$。

b. 证明：回归$\tilde{Y}_i=\beta_0+\beta_1X_i+v_i$满足重要概念4.3中的假设（假设对任意的$i$和$j$值，$w_i$独立于$Y_j$和$X_j$，并且具有有限的四阶矩）。

c. OLS估计量是一致的吗？

d. 能用通常的方法来构造置信区间吗？

e. 评价这句话：“变量X的测量误差是个严重的问题，而变量Y的测量误差则不是。”

7.3 研究妇女收人决定因素的劳动经济学家们发现了一个使人困惑的实证结果。选用随机选择的受雇佣妇女，他们用收人对妇女的孩子数和一系列的控制变量（年龄、学历、职业等）进行回归。他们发现在控制这些其他因素不变的情况下，有较多孩子的女性有较高的工资。解释样本选择如何有可能成为导致这种结果的原因。（提示：注意样本仅包括在职的妇女。）（这个实证难题激发了James Heckman对样本选择的研究，使他因此获得了2000年诺贝尔经济学奖）



附录　马萨诸塞州小学的考试数据

马萨诸塞州的数据是1998年公立小学学区的地区范围内的平均值。考试成绩取自于1998年春季马萨诸塞州公立学校中所有的四年级学生进行的马萨诸塞州综合评估系统(MCAS)考试。这个考试是由马萨诸塞州教育部门主办的,且对所有的公立学校都是强制性的。这里分析的数据是整体的总分数,它是考试中的英语、数学和科学部分的分数之和。

学生—教师比、享受午餐补助计划的学生的百分比和仍在学习英语的学生的百分比的数据,是1997—1998学年里每个小学的平均数,而且是从马萨诸塞州教育部门取得的。地区平均收入数据是从1990年美国人口普查数据中得到的。

原书空白页

第8章
面板数据回归

多元回归是控制变量影响的一种有力的工具,当然这里的变量我们需要掌握其数据。然而,如果得不到一些变量的数据,那么它们就不能包括在回归中,回归系数的 OLS 估计量就可能含有遗漏变量偏差。

本章描述了控制某些类型遗漏变量(不用实际观测它们)的一种方法。该方法要求一种特殊类型的数据,这种数据被称为面板数据。在面板数据中,每个观测单位或实体在两个或两个以上时期被观测。研究因变量随时间的变化,有可能消除那些在观测单位之间有差异但随时间不变的遗漏变量的影响。

本章中的实证应用的例子是关于醉酒驾车:酒税和醉酒驾车法对交通事故死亡率的影响是什么?我们用美国48个相邻的州1982年到1988年这7年间每一年的交通事故死亡率、酒税、醉酒驾车法以及有关的变量处理这个问题。这个面板数据集允许我们控制那些难以观测的变量,比如对饮酒和驾车的普遍态度。这些难以观测的变量在州之间不同但不随时间变化。面板数据也允许我们控制那些随时间变化但在各州之间不发生变化的变量,像新车安全措施的改善等。

8.1节描述了面板数据结构,并引入了醉酒驾车数据集。固定效应回归,作为面板数据回归分析的主要工具,是多元回归的一种推广,这种方法利用面板数据来控制那些在观测单位之间不同但随时间变化保持不变的变量。8.2节和8.3节介绍了固定效应回归,首先介绍了只有两个时期的情形,然后介绍了多个时期的情形。在8.4节,这些方法被扩展且合并了所谓的时间固定效应,这个时间固定效应控制了那些在观测单位之间不变但随时间变化的不可观测的变量。在8.5节,我们使用这些方法研究酒税和醉酒驾车法对交通事故死亡人数的影响。



8.1 面板数据

回想在1.3节中曾指出,面板数据(panel data,也称为纵向数据)是指 n 个不同实体在 T 个不同时期被观测的数据。本章所研究的州交通事故死亡率数据就是面板数据。这些数

据是 $n=48$ 个实体(州),其中每个实体被观测 $T=7$ 个时期(年份为 1982 年,……,1988 年),总共有 $7\times48=336$ 个观测值。

在描述截面数据时,使用下角标来表示实体是很有用的,例如 Y_i 是指第 i 个实体变量 Y。在描述面板数据时,我们需要一些额外的符号标明实体和时期,通过使用两个下角标而不是一个下角标就能够做到:第一个下角标 i 是指实体,第二个下角标 t 是指观测的时期。这样,Y_{it}就表示 n 个实体中第 i 个实体在 T 个时期中第 t 个时期被观测到的变量 Y。这个符号在重要概念 8.1 中作了总结。

重要概念 8.1

面板数据的符号表示

面板数据由同样的 n 个实体在两期或两期以上的观测值所组成。如果数据集包含变量 X 和 Y 的观测值,那么数据可表示为:

$$(X_{it},Y_{it}),i=1,2,\cdots,n,t=1,2,\cdots,T \tag{8.1}$$

其中,第一个下角标 i 指被观测的实体,第二个下角标 t 指被观测的时期。

与面板数据有关的一些额外术语描述了是否缺少一些观测值。均衡面板(balanced panel)含有所有的观测值,也就是说,变量对每个实体在每个时期都进行了观测。一个面板对至少一个实体在至少一个时期缺失数据,则该面板被称为非均衡面板(unbalanced panel)。由于交通事故死亡率数据集含有美国所有的 48 个州在整个 7 年的数据,因此它是均衡的。然而,如果有些数据缺少了(例如,如果我们没有某些州在 1983 年的死亡率数据),那么数据集就是非均衡的。本章所提到的方法是针对均衡面板来介绍的,不过,所有这些方法都能被用于非均衡面板,尽管实际上如何准确地去做取决于所使用的回归软件。

例子:交通死亡与酒税

在美国,每年大约有 40 000 例高速公路交通死亡事故。大约 1/3 的致命性撞车事故涉及司机醉酒驾车,而且这个比例在饮酒高峰期会上升。一项研究(Levitt 与 Porter,2000)表明,在凌晨 1 点到 3 点之间,行驶在公路上的司机有 25% 饮过酒,而饮酒的司机会造成致命撞车事故的可能性至少是没有饮酒司机的 13 倍。

在本章,我们研究了设计用来劝阻醉酒驾车的各种政府政策在减少交通死亡方面的实际效果如何。面板数据集包含了与交通死亡率和酒有关的变量,包括每个州每年交通事故的死亡人数、醉酒驾车法以及每个州的啤酒税。我们所使用的交通死亡测度是事故死亡率,它是在该州总人口中每 10 000 人的年度交通事故死亡人数。我们所使用的酒税测度是一箱啤酒的"实际"税,它是经通货膨胀调整转化为 1988 年美元[①]的啤酒税。在附录 8.1 中,更详细地介绍了这个数据。

图 8—1(a)是这些变量中的两个变量,即事故死亡率和每箱啤酒的实际税在 1982 年数据的散点图。这个散点图上的点表示某个给定的州 1982 年的事故死亡率和实际啤酒税。用事故死亡率对实际啤酒税回归,得到的 OLS 回归线也在该图中描绘出来,所估计的回归线是:



$$\widehat{FatalityRate} = \underset{(0.15)}{2.01} + \underset{(0.13)}{0.15}\, BeerTax \quad (1982\text{ 年数据}) \tag{8.2}$$

① 为了使税收在不同的时间里具有可比性,使用消费者价格指数(CPI)将它们转化为 1988 年美元。例如,由于通货膨胀,在 1982 年的 1 美元税收相应于以 1988 年美元计价的 1.23 美元的税收。

解，比较高的实际啤酒税与比较多而不是比较少的交通事故死亡率有联系。

我们是否应该得出这样的结论：增加啤酒税会导致更多的交通死亡事故？不一定，因为这些回归可能含有很大的遗漏变量偏差。许多因素影响死亡率，包括在每个州所驾驶的汽车质量，州高速公路是否处于良好的维护状态，大多数驾车是在城里还是在农村，公路上汽车的密度，醉酒驾车在社会上是否是可接受的。这些因素中的任何一个因素都可能与酒税相关，如果真是如此，那么它们就会导致遗漏变量偏差。处理这些遗漏变量偏差潜在来源的一种方法，就是搜集所有这些变量的数据，并将它们增加到公式(8.2)和公式(8.3)的年度截面回归中。不幸的是，有些变量，如醉酒驾车的文化接受度，可能很难测量，甚至不可能测量。

然而，如果在一个给定的州中，这些因素不随时间的变化而变化，那么可走另一个途径。由于我们有面板数据，因此即使我们不能够测量它们，我们也能使这些因素保持不变。为了这样做，我们使用固定效应的 OLS 回归。



8.2 两期面板数据："之前和之后"的比较

当获得每个州 $T=2$ 个时期的数据时，我们就可以把第二个时期的因变量值与第一个时期的因变量值进行比较。通过集中观测因变量的变化，这种"之前和之后"的比较实际上使那些难以观测的因素保持不变，那些难以观测的因素在各州之间是不同的，但在州内却不会随时间发生变化。

假设 Z_i 是决定第 i 个州交通事故死亡率的变量，它不随时间变化而变化（因此省略下角标 t）。例如，Z_i 可能是对醉酒驾车的地方文化态度，它变化缓慢，进而可以认为它在 1982 年到 1988 年之间是保持不变的。因此，将 Z_i 同实际啤酒税和交通事故死亡率联系起来的总体线性回归是：

$$FatalityRate_{it}=\beta_0+\beta_1 BeerTax_{it}+\beta_2 Z_i+u_{it} \tag{8.4}$$

其中，u_{it}是误差项，$i=1,\cdots,n,\ t=1,\cdots,T$。

由于 Z_i 不随时间变化而变化，因此，在公式(8.4)的回归模型中，它不会引起 1982 年到 1988 年之间交通事故死亡率的任何变化。因此在这个回归模型中，通过分析这两个时期之间的交通事故死亡率的变化可以除去 Z_i 的影响。为了在数学上理解这一点，对 1982 年和 1988 年这两年中的每一年，考虑方程(8.4)：

$$FatalityRate_{i1982}=\beta_0+\beta_1 BeerTax_{i1982}+\beta_2 Z_i+u_{i1982} \tag{8.5}$$

$$FatalityRate_{i1988}=\beta_0+\beta_1 BeerTax_{i1988}+\beta_2 Z_i+u_{i1988} \tag{8.6}$$

用公式(8.6)减去公式(8.5)，便可除去 Z_i 的影响：

$$FatalityRate_{i1988}-FatalityRate_{i1982}=\beta_1(BeerTax_{i1988}-BeerTax_{i1982})+(u_{i1988}-u_{i1982}) \tag{8.7}$$

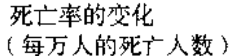
这种设定有个直观的解释。对醉酒驾车的文化态度影响一个州的醉酒驾车严重程度，进而影响一个州的交通事故死亡率。然而，如果在 1982 年和 1988 年之间它们没有发生变化，那么它们将不会引起这个州的事故死亡率的任何变化。更确切地说，交通事故死亡率随时间所发生的任何变化一定有其他的原因。在公式(8.7)中，这些其他的原因就是啤酒税的变化或误差项的变化（它反映了决定交通事故死亡率变化的其他因素的变化）。

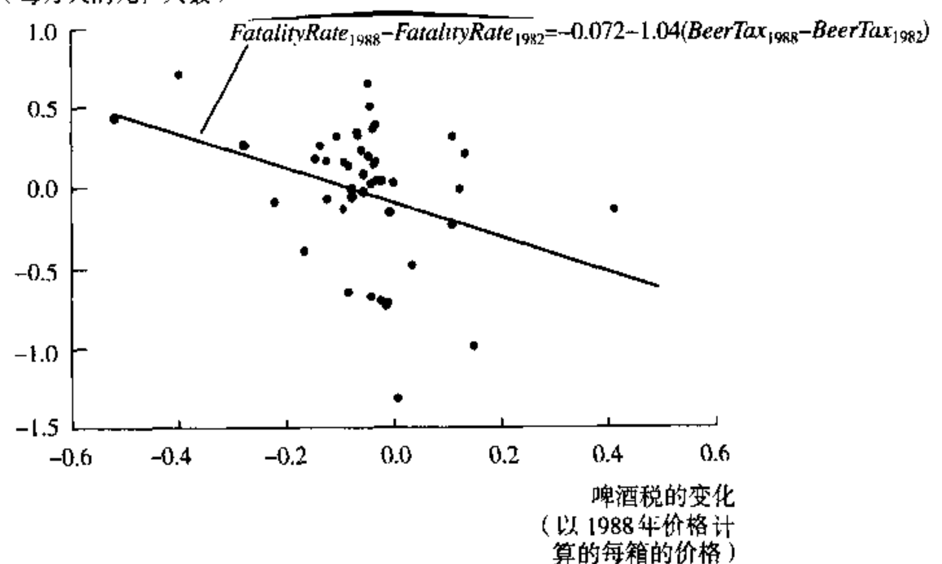
对公式(8.7)中变化的回归设定，除去了不随时间发生变化的且难以观测的变量 Z_i 的影响。换句话说，分析 Y 和 X 的变化具有控制随时间保持不变的变量的作用，这样就除去了这个遗漏变量偏差的来源。

图 8—2 表示了我们的数据集中,48 个州在 1982 年到 1988 年之间交通事故死亡率变化对实际啤酒税变化的散点图。图 8—2 中的点表示给定的州在 1982 年到 1988 年之间交通事故死亡率的变化和实际啤酒税的变化。使用这些数据所估计的并在图中画出的 OLS 回归线是:

$$\widehat{FatalityRate_{i1988} - FatalityRate_{i1982}} = \underset{(0.065)}{-0.072} - \underset{(0.36)}{1.04}\,(BeerTax_{i1988} - BeerTax_{i1982}) \tag{8.8}$$

其中,所包含的截距项考虑到了在实际啤酒税没有发生变化的情况下,交通事故死亡率平均变化的可能性是非零的。





注:这是 48 个州在 1982 年到 1988 年之间交通事故死亡率变化和实际啤酒税变化的散点图。在交通事故死亡率变化和啤酒税变化之间存在负向关系。

图 8—2 交通事故死亡率和啤酒税的变化(1982—1988 年)

与截面回归结果相比较,上式中实际啤酒税变化的估计效应是负的,正如经济理论所预测的那样。总体斜率系数为零的假设在 5% 的显著性水平下被拒绝。根据这个估计系数,每箱啤酒实际税增加 1 美元会使每 10 000 人中交通事故死亡率减少 1.04 例。这个估计效应是很大的:在这些数据中平均死亡率大约是 2(也就是说,总体中每 10 000 人中每年有 2 例死亡事故),因此,这个估计值表明每箱啤酒实际税只增加 1 美元就能使交通事故死亡率减少一半。

通过研究交通事故死亡率随时间发生的变化,公式(8.8)中的回归控制了诸如对醉酒驾车的文化态度这样的固定因素,但是存在许多影响交通安全的因素,如果它们随时间发生变化,并且与实际啤酒税相关,那么忽略它们将会引起遗漏变量偏差。在 8.5 节中,我们进行了更细致的分析,那个分析控制了好几个这样的因素,所以目前最好暂不给出任何关于实际啤酒税对交通事故死亡率影响的实质性结论。

当数据在两个不同年份里被观测时,可以采用这种“之前和之后”的回归分析方法。然而,我们的数据集包括了 7 个不同年份的观测值,而丢弃那些潜在有用的额外数据似乎是很愚蠢的。但是,当 $T>2$ 时,这种“之前和之后”的方法不能被直接应用。为了分析我们的面板数据集中的所有观测值,我们使用固定效应回归方法。



8.3 固定效应回归

固定效应回归是控制面板数据中遗漏变量的一种方法,这里的遗漏变量在实体(州)之间发生变化但不随时间发生变化。不像8.2节中的"之前和之后"比较,当每个实体有两个或多个时期的观测值时,可以使用固定效应回归。

固定效应回归模型有 n 个不同的截距,每个截距对应于一个实体。这些截距能被表示为一组二元变量(或指示变量)。这些二元变量吸收了那些在州之间不同但又不随时间变化的所有遗漏变量的影响。

8.3.1 固定效应回归模型

考虑将公式(8.4)中的因变量(*FatalityRate*)和观测到的回归因子(*BeerTax*)分别表示为 Y_{it} 和 X_{it} 的回归模型:

$$Y_{it}=\beta_0+\beta_1 X_{it}+\beta_2 Z_i+u_{it} \tag{8.9}$$

其中,Z_i 是那些在州之间发生变化但又不随时间变化而变化的难以观测的变量(例如,Z_i 表示对醉酒驾车的文化态度)。我们想在保持那些难以观测的州的特征 Z 不变的情况下,估计 X 对 Y 的影响 β_1。

由于 Z_i 在州之间发生变化,但又不随时间变化而变化,因此公式(8.9)中的总体回归模型可以被解释为含有 n 个截距,即每个州对应一个截距。具体地说,假定 $\alpha_i=\beta_0+\beta_2 Z_i$,那么公式(8.9)变为:

$$Y_{it}=\beta_1 X_{it}+\alpha_i+u_{it} \tag{8.10}$$

公式(8.10)就是**固定效应回归模型**(fixed effects regression model),其中将 $\alpha_1,\cdots,\alpha_n$ 看做为待估计的未知截距,每个 α 对应着一个州。α_i 可以被解释为来自公式(8.10)中考虑第 i 个州的总体回归线因州而异的截距,这个总体回归线就是 $\alpha_i+\beta_1 X_{it}$。总体回归线的斜率系数 β_1 对所有的州来说都是一样的,但总体回归线的截距在各州之间却发生变化。因此,截距变化的根源是变量 Z_i,它(Z_i)在各州之间发生变化但又不随时间而变化。

固定效应回归模型中因州而异的截距项,也可以利用表示单个州的二元变量来表达。6.3节考虑了这种情形,在那里观测值属于两个组中的一组,且对两个组而言,总体回归线有相同的斜率,但有不同的截距(见图6—8(a))。那个总体回归线在数学上是使用表示两组之一的单个二元变量来表达的(见重要概念6.4中的第1种情况)。如果在我们的数据集中只有两个州,那么二元变量回归模型在这里将会适用。然而,由于我们不止有两个州,所以我们需要增加额外的二元变量来捕捉公式(8.10)中所有因州而异的截距。

为了使用二元变量展开固定效应回归模型,假定 $D1_i$ 是个二元变量,当 $i=1$ 时,$D1_i=1$;否则,$D1_i=0$。假设当 $i=2$ 时,$D2_i=1$;否则,$D2_i=0$;如此等等。我们不能够把所有的 n 个二元变量和一个共同的截距项都包括进来,因为如果那样的话,回归因子将是完全多重共线的(见练习8.2)。因此,我们任意地省略第一组二元变量 $D1_i$,从而,公式(8.10)中的固定效应回归模型能被等价地写为:

$$Y_{it}=\beta_0+\beta_1 X_{it}+\gamma_2 D2_i+\gamma_3 D3_i+\cdots+\gamma_n Dn_i+u_{it} \tag{8.11}$$

其中,$\beta_0,\beta_1,\gamma_2,\cdots,\gamma_n$ 是待估计的未知系数。为了导出公式(8.11)中的系数和公式(8.10)中的截距之间的关系,比较这两个方程所对应的每个州的总体回归线。在公式(8.11)中,第一个州的总体回归方程是 $\beta_0+\beta_1 X_{it}$,所以 $\alpha_1=\beta_0$。对第二个州和其余各州而言,总体回

归方程是$\beta_0+\beta_1X_{it}+\gamma_i$，所以对于$i\geq2$，有$\alpha_i=\beta_0+\gamma_i$。这样，就有了表达固定效应回归模型的两个等价的方法，即公式(8.10)和公式(8.11)。在公式(8.10)中，它是根据n个因州而异的截距来表示的。在公式(8.11)中，固定效应模型有一个共同的截距和$n-1$个二元回归因子。在这两个表达式中，X的斜率系数在不同的州之间是相同的。公式(8.10)中因州而异的截距和公式(8.11)中的二元回归因子具有相同的来源，即在各州间发生变化但又不随时间变化的难以观测的变量Z_i。

重要概念 8.2

固定效应回归模型

固定效应回归模型是：

$$Y_{it}=\beta_1X_{1,it}+\cdots+\beta_kX_{k,it}+\alpha_i+u_{it} \tag{8.12}$$

其中，$i=1,2,\cdots,n$，$t=1,2,\cdots,T$，$X_{1,it}$是实体i在时期t的第一个回归因子值，$X_{2,it}$是第二个回归因子值，如此等等，而$\alpha_1,\cdots,\alpha_n$是特定实体的截距。

同样，固定效应回归模型也可以写成一个共同的截距、X变量以及除了一个实体以外的代表其余实体的所有$n-1$个二元变量：

$$Y_{it}=\beta_1X_{1,it}+\cdots+\beta_kX_{k,it}+\gamma_2D2_i+\gamma_3D3_i+\cdots+\gamma_nDn_i+u_{it} \tag{8.13}$$

其中，如果$i=2$，那么$D2_i=1$；否则，$D2_i=0$，依此类推。

推广到多元变量X。如果存在决定Y变化的其他可观测的因素，而且如果这些因素与变量X相关并随时间变化而变化，那么这些决定性因素也应该被包括在回归中以避免发生遗漏变量偏差，由此便产生了具有多个回归因子的固定效应模型，被总结在重要概念8.2中。

固定效应回归模型的最小二乘假设。对固定效应回归模型而言，存在五个最小二乘假设：在重要概念5.4中多元回归模型的四个假设(对面板数据进行了修正)加上第五个新的假设。在截面数据中，误差项在实体之间是不相关的，以回归因子为条件。把这个假设推广到面板数据中，便得到了第五个假设，这个假设假定，误差项不仅在实体之间不相关，在不同时期也不相关，也以回归因子为条件。这些假设在概念上类似于多元回归模型的最小二乘假设，但是由于和面板数据集有关的符号十分复杂，因此它们的数学表达也相当复杂。附录8.2中叙述并讨论了这些假设。

8.3.2 估计和推断

从原则上说，固定效应回归模型(公式(8.13))的二元变量设定也可以使用OLS来估计。然而，这个回归有$k+n$个回归因子(k个X，$n-1$个二元变量和一个截距)，因此，如果实体数很大，那么在实际中，这个OLS回归是冗长的，在一些软件包里甚至不可能运行。因此，经济计量软件中一般含有固定效应回归模型的OLS估计的特定程序。这些特定的程序等价于对全部二元变量回归使用OLS，但是由于程序运用了固定效应回归在代数上所出现的一些数学简化，因此程序运行比较快。

"去均值实体(entity-demeaned)"的OLS算法。回归软件用两个步骤来计算OLS固定效应的估计量。在第一步中，每个变量减去每个对应实体的均值。在第二步中，使用"去均值实体"变量来估计回归。具体来说，考虑公式(8.10)中固定效应模型中单个回归因子的情形，并对公式(8.10)两边(关于时间t)取平均值，那么就有$\overline{Y}_i=\beta_1\overline{X}_i+\alpha_i+\overline{u}_i$，其中$\overline{Y}_i=$



8.4 带有时间固定效应的回归

就像每个实体的固定效应能够控制那些不随时间变化而变化但在各实体间却不相同的变量一样，时间固定效应也能够控制在实体间不变但随时间变化的变量。

由于新车安全设施的改善在全国范围内采用，因此它们有助于减少所有州的交通死亡事故。因而，将汽车安全作为一个随时间变化但对所有的州有相同值的遗漏变量似乎是合理的。我们用 S_t 表示汽车安全，公式(8.9)中的总体回归可被修改为包括汽车安全效应的方程：

$$Y_{it}=\beta_0+\beta_1X_{it}+\beta_2Z_i+\beta_3S_t+u_{it} \tag{8.16}$$

其中，S_t 是难以观测的，这里单个下角标"t"强调安全性随时间发生变化但在各州间不变。由于 β_3S_t 代表了决定 Y_{it} 的变量，因此如果 S_t 与 X_{it} 相关，那么回归中忽略 S_t 将会导致遗漏变量偏差。



8.4.1 只有时间效应

目前，假设变量 Z_i 不存在，所以 β_2Z_i 项可以从公式(8.16)中省略掉，尽管 β_3S_t 项留下来了。我们的目标是控制 S_t，估计 β_1。

虽然 S_t 是难以观测的，但是由于它随时间发生变化但在州之间不发生变化，因此，就像除去在州之间发生变化但不随时间变化的 Z_i 的影响一样，除去它的影响也是可能的。也就是说，可以用一组 T 个二元变量代替 β_3S_t，每个二元变量表示一个不同的年份。具体来说，假设如果 t 是样本中的第一个时期，那么 $B1_t=1$；否则，$B1_t=0$。假设如果 t 是样本中的第二个时期，$B2_t=1$；否则，$B2_t=0$，依此类推。这些二元变量 $B1_t,\cdots,BT_t$ 被称为时间效应(time effects)。

含有单个 X 回归因子和 $T-1$ 个时间效应的时间效应回归模型(time effects regression model)为：

$$Y_{it}=\beta_0+\beta_1X_{it}+\delta_2B2_t+\cdots+\delta_TBT_t+u_{it} \tag{8.17}$$

其中，$\delta_2,\cdots,\delta_T$ 是未知系数。同公式(8.11)中的固定效应回归模型一样，这个时间效应模型中也包含了截距，而且为了防止完全多重共线性，第一个二元变量($B1_t$)被省略了。

在交通事故死亡率一例的回归中，公式(8.17)中的时间固定效应设定允许我们消除像在全国范围内引进汽车安全标准这样的遗漏变量所引起的偏差，在给定的年份里，汽车安全标准随时间发生变化但在各州间不变。

8.4.2 时间和州固定效应

如果一些遗漏变量不随时间变化而变化但在州间却发生变化(如文化规范)，而其他一些遗漏变量在州之间不变化但随时间变化而变化(如全国性的安全标准)，那么把州效应和时间效应都包括进来是合适的，通过在回归中把 $n-1$ 个州的二元变量和 $T-1$ 个时间二元变量以及一个截距项包括进来，就可以做到这一点。合并的时间和实体固定效应回归模型(time and entity fixed effects regression model)为：



$$Y_{it}=\beta_0+\beta_1X_{it}+\gamma_2D2_i+\cdots+\gamma_nDn_i+\delta_2B2_t+\cdots+\delta_TBT_t+u_{it} \tag{8.18}$$

其中，$\beta_0,\beta_1\gamma_2,\cdots,\gamma_n,\delta_2,\cdots,\delta_T$ 都是未知系数。

这个合并的时间和州固定效应回归模型，消除了那些不随时间发生变化的难以观测的

变量和那些在州之间不发生变化的难以观测的变量所引起的遗漏变量偏差。

当存在其他的可观测的回归因子“X”时，这些回归因子也应该出现在公式(8.18)中。

估计。时间固定效应模型和时间与实体(州)固定效应模型都是多元回归模型的变体。这样，它们的系数可以通过包括额外的时间二元变量用 OLS 来估计。一些软件并入了计算合并的时间固定效应与实体固定效应回归的算法。这个算法在计算上比公式(8.18)中完全的二元变量模型的 OLS 估计更有效。

在交通死亡例子中的应用。将时间效应增加到州固定效应回归中就生成了回归线的 OLS 估计：

$$\widehat{FatalityRate} = \underset{(0.25)}{-0.64}\ BeerTax + StateFixedEffects + TimeFixedEffects \tag{8.19}$$

这种模型设定包括啤酒税、47个州的二元变量(州固定效应)、6年的二元变量(时间固定效应)和一个截距，因此实际上这个回归右边有 1 + 47 + 6 + 1 = 55 个变量！由于时间和州二元变量系数与截距不是我们主要感兴趣的对象，因此这里没有给出。

把时间效应包括进来对实际啤酒税和交通事故死亡率之间的估计关系几乎没有影响(比较公式(8.15)和公式(8.19))。实际啤酒税系数在5%的水平下仍旧是显著的($t = -0.64/0.25 = -2.56$)。

在实际啤酒税和交通事故死亡率之间的这个估计关系，不受来自于不随时间或州的变化而变化的变量的遗漏变量偏差的影响。然而，影响交通死亡事故的许多重要的决定性因素并没有被包含在这个类别里，因此，这种设定仍旧受遗漏变量偏差的影响。有了固定效应回归的工具，我们可对手中这些数据进行更全面的实证分析。

8.5 醉酒驾车法与交通事故死亡率

酒税只是规劝醉酒驾车的一种方法。每个州对醉酒驾车的惩罚不同，而且制裁醉酒驾车的州除了提高税收以外，还可以通过加强立法来实现对醉酒驾车的全面制裁。如果是这样的话，忽略这些法律制度会在实际啤酒税对交通事故死亡率影响的 OLS 估计量中，甚至在含有州和时间固定效应的回归中，产生遗漏变量偏差。此外，由于车辆使用部分依赖于司机是否有工作，而且由于税收的变化能够反映经济条件(州预算赤字会导致税收上升)，因此忽略州经济条件也会导致遗漏变量偏差。

本节中，在经济条件保持不变的情况下，我们将前面的分析推广到研究饮酒法(包括啤酒税)对交通事故死亡率的影响。通过把代表其他醉酒驾车法和州经济条件的回归因子包括在面板数据回归中来进行估计，我们就可以实现这个目标。

计算结果归纳在表8—1中。表8—1的格式和第5、第6、第7章中的回归结果表的格式相同：每列报告不同的回归结果，而每行则报告系数估计值、标准误，F 统计量、p 值或者有关回归的其他信息。

表8—1中第(1)列给出了在没有州和时间固定效应的情况下，交通事故死亡率对实际啤酒税的 OLS 回归的结果。同1982年和1988年的截面回归(公式(8.2)和公式(8.3))一样，实际啤酒税的系数是正的(0.36)，而且第(1)列的估计值在5%的水平下在统计上显著地异于0。根据这个估计值，提高啤酒税会增加交通事故死亡率！然而，包括州固定效应的第(2)列中的回归(和前面公式(8.15)所给出的一样)表明，回归(1)中正的系数是由遗漏变量偏差造成的(实际啤酒税系数是 −0.66)。当包括固定效应时，回归的 $\bar{R}^2$ 从0.090跳到



0.889，很显然，州固定效应解释了数据变差的大部分变化。

表 8—1　　醉酒驾车法对交通死亡影响的回归分析

因变量：交通事故死亡率（每 10 000 人口的死亡人数）						
回归因子	(1)	(2)	(3)	(4)	(5)	(6)
啤酒税	0.36**	-0.66**	-0.64*	-0.45*	-0.70**	-0.46*
	(0.05)	(0.20)	(0.25)	(0.22)	(0.25)	(0.22)
饮酒年龄为 18				0.028	-0.011	
				(0.066)	(0.064)	
饮酒年龄为 19				-0.019	-0.078	
				(0.040)	(0.049)	
饮酒年龄为 20				0.031	-0.102*	
				(0.046)	(0.046)	
饮酒年龄						-0.002
						(0.017)
强制性坐牢				0.013	-0.026	
				(0.032)	(0.065)	
强制性社区服务				0.033	0.147	
				(0.115)	(0.137)	
强制性坐牢或强制性社区服务						0.031
						(0.076)
每位驾驶员的平均车辆里程				0.008	0.017	0.009
				(0.008)	(0.010)	(0.008)
失业率				-0.063**		-0.063**
				(0.012)		(0.012)
人均实际收入（对数）				1.81**		1.79**
				(0.47)		(0.45)
州效应	否	是	是	是	是	是
时间效应	否	否	是	是	是	是
检验排除变量组的 F 统计量和 p 值						
时间效应 =0			2.47	11.44	2.28	11.59
			(0.024)	(<0.001)	(0.037)	(<0.001)
饮酒年龄系数 =0				0.48	2.09	
				(0.696)	(0.102)	
坐牢、社区服务系数 =0				0.17	0.59	
				(0.845)	(0.557)	
失业率、人均收入系数 =0				38.29		40.12
				(<0.001)		(<0.001)
$\bar{R}^2$	0.090	0.889	0.891	0.926	0.893	0.926

注：这些回归是使用附录 8.1 中所介绍的美国 48 个州从 1982 年到 1988 年（总共 336 个观测值）的面板数据来估计的。标准误在系数下面的括号中给出，p 值在 F 统计量下面的括号中给出。个别系数在 *5% 的显著性水平下或 **1% 的显著性水平下在统计上是显著的。



归,研究了该分析结果对使用饮酒年龄(用饮酒年龄本身代替三个指示变量)的不同函数形式以及对合并两个二元惩罚变量的敏感性。回归(4)中的结果对这些变化是不敏感的。

这种分析的优势是,把州和时间固定效应包括进来,缓和了那些不随时间发生变化(像对醉酒驾车的文化态度)或者在州间不发生变化(像车辆的安全革新措施)的难以观测的变量所引起的遗漏变量偏差的威胁。然而像通常一样,认识到这种分析可能的局限性也是非常重要的。遗漏变量偏差的一个潜在的来源是,这里所使用的酒税的测度,即实际啤酒税,可能会和其他酒税一起变化,这表明我们在解释结果时不能单单局限于啤酒税。一个比较细微的可能性是实际啤酒税的增加可能与公共教育运动有关,这也许是对政治压力的反应。如果是这样的话,实际啤酒税的变化可以因为这一广泛的运动而扩大对醉酒驾车的效应。

这些结果为控制醉酒驾车和交通事故死亡率提供了一个富有挑战性的测度建议。根据这些估计值,严厉的惩罚和提高最低合法饮酒年龄对交通事故死亡率都没有重要影响。相反,有证据表明增加酒税,就像用实际啤酒税所估计的那样,确实减少了交通死亡人数。然而,这个效应的大小估计得还不够精确。[①]

8.6 结论

本章说明了如何使用相同实体随时间变化的多个观测值来控制实体间不同但又不随时间变化的难以观测的遗漏变量。关键的要点是,如果难以观测的变量不随时间发生变化,那么因变量的任何变化一定是由于除这些固定特征以外的其他因素的影响。如果对醉酒驾车的文化态度在一个州内7年间不发生明显变化,那么在这7年间对交通事故死亡率变化的解释一定依赖于别的因素。

为了拓展这个观点,你需要同一个实体在两期或者两期以上的时间内所观测到的数据,也就是说,你需要面板数据。有了面板数据,第2部分的多元回归模型可以被推广到包括二元变量的全集,每个二元变量对应着一个实体,这就是固定效应回归模型,它能够用OLS进行估计。处理固定效应回归模型的方法就包括时间固定效应,它控制了随时间变化但在实体间不变的难以观测的变量。实体和时间固定效应都可以被包含在回归中,以控制那些在实体间变化但不随时间变化的变量和那些随时间变化但在实体间不发生变化的变量。

虽然有这些优点,但是实体和时间固定效应回归却不能控制在实体间变化且随时间变化的遗漏变量。很显然,面板数据方法需要面板数据,但面板数据通常是得不到的。这样,当面板数据方法完成不了任务时,还需要另一种能够消除那些难以观测的遗漏变量影响的方法。完成这项任务的一个有力而又一般的方法就是工具变量回归,这是第10章的主题。

总结

1. 面板数据是由多个(n)实体——州、企业、人等——的观测值构成的,其中每个实体都被观测两期或两期以上(T)。

2. 带有实体固定效应的回归,用来控制那些在实体间不同但又不随时间发生变化的难以观测的变量。



3. 当有两个时期时,固定效应回归可以用因X的变化而发生的Y从第一期到第二期变

① 如果你有兴趣了解这些数据的进一步分析,请参见Ruhm(1996)。如果你对更多地了解醉酒驾车和酒,以及更一般地对酒经济学感兴趣,请见Cook与Moore(2000)。

化的“之前和之后”回归进行估计。

4. 实体固定效应回归,可以采用在方程中包含 $n-1$ 个实体的二元变量、可观测的自变量(X)和一个截距项的方法进行估计。

5. 时间固定效应控制了那些在实体间相同但随时间发生变化的难以观测的变量。

6. 包含时间和实体固定效应的回归,可以采用在方程中包含 $n-1$ 个实体的二元变量和 $T-1$ 个时期的二元变量,加上自变量 X 和一个截距项的方法进行估计。

重要术语

面板数据　均衡面板　非均衡面板　固定效应回归模型　时间固定效应　时间效应回归模型　时间和实体固定效应回归模型

复习概念

8.1　为什么必须使用两个下角标 i 和 t 来描述面板数据? i 指的是什么? t 又指的是什么?

8.2　一位研究人员正使用一个面板数据集,其中含有 $n=1\ 000$ 名工人、$T=10$ 年(从1991年到2000年)的面板数据,变量包括工人收入、性别、教育和年龄。该研究人员拟研究教育对收入的影响,请给出一些与教育和收入变化相关的且难以观测的因人而异的变量的例子。你能想象出那些可能与教育和收入都相关的因时间变化而变化的变量的例子吗? 在面板数据回归中,你如何控制这些因人而异和因时间而异的效应呢?

8.3　你在回答问题8.2时所提出的回归,能否被用于估计性别对个人收入的影响? 该回归能否被用于估计全国失业率对个人收入的影响? 请解释原因。

练习

8.1　这个问题涉及表8—1中所归纳的醉酒驾车面板数据回归。

*a. 新泽西州的人口是8 100 000。假设新泽西州每箱啤酒增加1美元税收(以1988年美元计算),使用第(4)列中的结果预测下一年将被挽救的生命数。为你的答案构造一个95%的置信区间。

b. 新泽西州的饮酒年龄是21岁。假设新泽西州将其饮酒年龄降低到18岁,使用第(4)列中的结果预测下一年中交通事故死亡率的变化。为你的答案构造一个95%的置信区间。

*c. 假设新泽西州下一年的实际人均收入增加1%,使用第(4)列中的结果预测下一年中交通事故死亡率的变化。为你的答案构造一个95%的置信区间。

d. 该回归是否应该包括时间效应? 为什么?

*e. 第(5)列中啤酒税系数的估计值在1%的水平下是显著的,第(4)列中的估计值在5%的水平下是显著的,这是否意味着第(5)列中的估计值更可靠?

f. 一位研究人员推测,失业率对西部地区州的交通事故死亡率的影响不同于其他州。你怎样检验这个假设?(提示:要明确你所使用的回归设定和统计检验)

8.2　考虑方程(8.11)中的固定效应模型的二元变量形式,这里只是多加了一个额外

回归因子 $D1_t$，即假定：

$$Y_{it}=\beta_0+\beta_1X_{it}+\gamma_1D1_t+\gamma_2D2_t+\cdots+\gamma_nDn_t+u_{it} \tag{8.20}$$

a. 假设 $n=3$，证明：二元变量回归因子和"常数"回归因子是完全多重共线的，也就是说，将 $D1_t$，$D2_t$，$D3_t$ 和 $X_{0,it}$ 中任何一个变量表示成其他变量的完全线性函数，其中，对于所有的 i 和 t 都有 $X_{0,it}=1$。

b. 对于一般的 n，证明(a)中的结论。

c. 如果你尝试用 OLS 估计公式(8.20)中的回归系数，那么将会发生什么？

8.3　7.2 节中列出了对回归研究中的内部有效性的五个潜在威胁。将这些威胁因素应用于 8.5 节中的经验分析，从而得出关于它的内部有效性的结论。

附录 8.1　州交通事故死亡率数据集

该数据是美国"下 48 个州"(除了阿拉斯加和夏威夷)从 1982 年到 1988 年的年度数据。交通事故死亡率是指，在给定的年份里生活在给定的州内的每 10 000 人中交通事故的死亡人数。交通事故死亡率数据是从美国交通部重大事故报告系统中获得的。啤酒税是每箱啤酒的税收，它是对一个州中酒税的一般性测度。表 8—1 中的饮酒年龄变量是表示合法饮酒年龄是否为 18 岁、19 岁或 20 岁的二元变量。表 8—1 中的两个二元惩罚变量描述了该州对醉酒驾车初犯定罪的最低审判要求：如果按该州的标准要求坐牢，那么"强制性坐牢"等于 1；否则，等于 0。如果该州要求做社区服务，那么"强制性社区服务"等于 1；否则，等于 0。州的年度车辆行驶总里程数据从美国交通部得到，个人收入数据从美国经济分析局获得，而失业率数据则从美国劳动统计署获得。

很感谢北卡罗莱纳大学经济系 Christopher J. Ruhm 教授向我们提供了这些数据。

附录 8.2　固定效应回归的假设

重要概念 8.2 中列出了固定效应回归模型的五个最小二乘假设。就单个回归因子而言，这五个假设是：

1. $E(u_{it}|X_{i1},X_{i2},\cdots,X_{iT},\alpha_i)=0$；

2. $(X_{i1},X_{i2},\cdots,X_{iT},Y_{i1},Y_{i2},\cdots,Y_{iT})$，$i=1,2,\cdots,n$，取自于它们联合分布的独立同分布的样本；

3. (X_{it},u_{it})有非零的有限的四阶矩；

4. 不存在完全多重共线性；

5. 对于 $t\neq s$，有 $\text{cov}(u_{it},u_{is}|X_{it},X_{is},\alpha_i)=0$。

对多个回归因子而言，应该用全列 $X_{1,it},X_{2,it},\cdots,X_{k,it}$代替 X_{it}。

第一个假设是指，在给定回归因子的条件下，误差项的条件均值为零。这与重要概念 5.4 中的第一个最小二乘假设相同，被推广到包括二元回归因子和第 i 个实体随时间变化的关于 X 的所有 T 个观测值，在第 2 部分中这个假设的讨论可以被直接推广。



第二个假设将多元回归的独立同分布假设推广到面板数据。如果实体是从总体中通过简单随机抽样来选择的，那么这个假设就成立。因而，一个实体的变量除了与另一个实体变量同分布外，还独立于另一个实体变量的分布，也就是说，对于 $i=1,2,\cdots,n$，变量是独立同分布的。迄今为止，这种推理与截面数据推理一样。然而，在面板数据中，实体按照时间排

第 9 章

二元因变量回归

除种族不同外，两个条件完全相同的人走进一家银行申请一笔很大数额的抵押贷款，目的是每人够买一套房子，两套房子的条件也完全相同。银行是否会以同样的方式对待他们？他们是否有同等可能性让他们的抵押贷款申请被接受？根据法律，他们必须受到同等对待，但是他们是否能确实得到同等对待是银行监管者很关心的一个问题。

有许多合理的理由批准贷款和拒绝贷款。例如，如果所提出的贷款偿付额占用了申请者的大部分或者全部的月收入，那么信贷员可以有理由地拒绝这笔贷款。毕竟信贷员也是人，他们也会犯普通的错误，所以对一个少数民族申请者拒绝的例子无法证明任何关于种族歧视问题的存在。因而，许多种族歧视的研究寻找关于种族歧视的统计证据，也就是说，从大量的数据集中寻找能够证明白人和少数民族是否被区别对待的证据。

但是在抵押贷款市场上，人们该如何精确地检查种族歧视的统计证据？一个出发点就是，比较少数民族申请者和白人申请者抵押贷款被拒绝的比重。在本章所研究的数据中，搜集了 1990 年在马萨诸塞州波士顿地区的抵押贷款申请，资料显示有 28% 的黑人申请者被拒绝，但仅有 9% 的白人申请者被拒绝。但是，这个比较并没有真正回答本章开头所提出的问题，因为黑人申请者和白人申请者不一定“除了他们的种族外都是相同的”。相反，我们需要一种在保持其他申请者特征不变的情况下比较拒绝率的方法。

这看起来像是多元回归分析的事情——它确实是，除了在方法上。这个方法就是因变量——申请者是否被拒绝——是二元变量的回归方法。在第 2 部分中，我们经常将二元变量用做回归因子，它们没有引起任何特别的问题。但是当因变量是二元变量时，事情就变得困难了：用一条直线去拟合只取 0 和 1 这两个值的因变量意味着什么？

这个问题的答案就是把这个回归函数解释为一个被预测的概率。在 9.1 节中我们讨论了这个解释，它允许我们将第 2 部分中的多元回归模型应用于二元因变量。9.1 节复习了这个“线性概率模型”。但是，对这个被预测的概率的解释表明，其他形式的非线性回归模型对模拟这些概率可能效果会更好。在 9.2 节中我们讨论了这些被称为“probit”回归和“logit”回归的方法。9.3 节论述了用来估计“probit”回归和“logit”回归系数的方法，即极大似然估计方法。这节的内容可以是选择性的。在 9.4 节中，我们将这些方法应用于波士顿

抵押贷款申请的数据集中，看一看在抵押贷款市场中是否存在种族歧视的证据。

本章所考虑的二元因变量，是个具有有限范围的因变量的例子，换句话说，它是个受限因变量(limited dependent variable)。对于其他类型的受限因变量，例如取多个离散值的因变量，我们在附录9.3中做了讨论。

9.1 二元因变量与线性概率模型

不论抵押贷款申请被接受还是被拒绝，都是二元变量的一个例子。许多其他重要的问题也涉及二元结果。奖学金对某个人上大学的决策的影响是什么？是什么因素决定青少年是否吸烟？是什么因素决定一国是否能得到外国援助？是什么因素决定工作申请是否成功？在所有这些例子中，我们所关心的结果都是二元的：学生上大学还是不上大学；青少年吸烟还是不吸烟；一国得到外国援助还是没有得到外国援助，申请者获得了工作还是没有获得工作。

本节论述了二元因变量回归和连续因变量回归之间的区别，然后介绍一种含有二元因变量的最简单的回归模型——线性概率模型。

9.1.1 二元因变量

本章所研究的应用问题是，种族是否是拒绝一项抵押贷款申请的一个因素，相应的二元因变量是，抵押贷款申请是否被拒绝。所用的数据是波士顿联邦储备银行的研究人员在住房抵押披露法(HMDA)的要求下编辑的一个较大数据集中的一个子集，而且数据与1990年在马萨诸塞州波士顿地区所填写的抵押贷款申请档案有关。附录9.1中描述了波士顿HMDA数据。

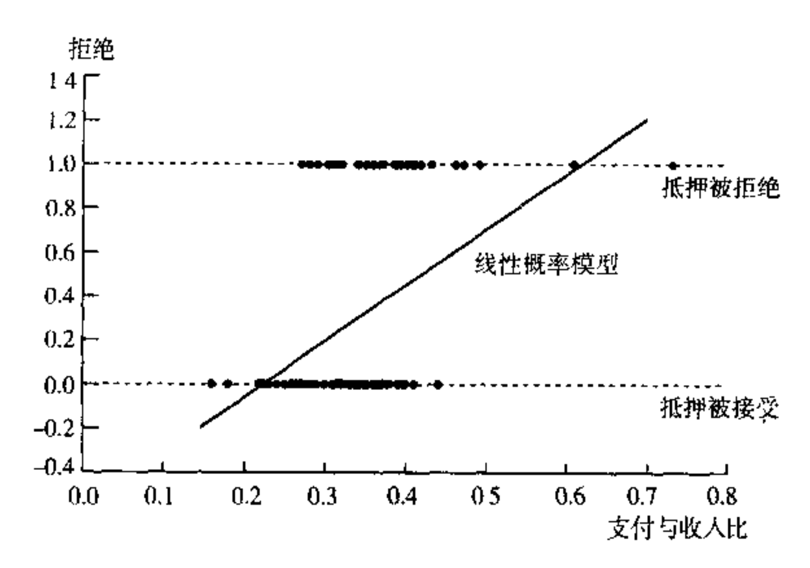
抵押贷款的申请是很复杂的，银行信贷员的决定过程也是如此。一个信贷员必须能预测该申请者是否有能力支付他(或她)的贷款，一条重要的信息就是要求的贷款支付额相对于申请者收入的大小。任何贷款人都知道，支付一项相当于你月收入10%的贷款要比支付相当于你月收入50%的贷款更容易。因此，我们首先看两个变量之间的关系：二元因变量*deny*(如果抵押贷款申请被拒绝，那么*deny*等于1；否则，*deny*等于0)，以及连续变量*P/I ratio*(支付与收入的比)，它是申请者的预期月贷款支付总额与他(或她)的月收入之比。

图9—1给出了*deny*和*P/I ratio*之间关系的散点图，数据是从2 380个观察值中截取的127个(用数据集的这个子集更容易理解这个散点图)。由于变量*deny*是二元变量，因此这个散点图看起来不同于第2部分的散点图。虽然如此，它看上去仍然显示了*deny*和*P/I ratio*之间的关系：*P/I ratio*低于0.3的申请者很少被拒绝，但是*P/I ratio*超过0.4的大部分申请者被拒绝了。

*P/I ratio*和*deny*之间的这种正相关关系(*P/I ratio*越高，拒绝比例越大)连同使用这127个观察值所估计出的OLS回归线，一起反映在图9—1中。和以前一样，这条回归线描绘的是把*deny*作为*P/I ratio*这一回归因子的函数的预测值。例如，当*P/I ratio* = 0.3时，*deny*的预测值是0.20。但是，对二元变量*deny*的预测值为0.20而言，如何解释这个数值的含义呢？



回答这个问题的关键(这也是理解二元因变量回归的一般性问题)，是把该回归关系解释为建立一个因变量等于1的概率模型。这样，预测值0.20就可被解释为，当*P/I ratio* = 0.3时，估计出的被拒绝的概率为20%，换句话说，如果有许多*P/I ratio* = 0.3的申请，那么



注:具有高的债务支付与收入比(*P/I ratio*)更可能使抵押贷款申请者的申请被拒绝(如果被拒绝,那么 *deny* = 1;否则,*deny* = 0)。线性概率模型使用直线建立以 *P/I ratio* 为条件的被拒绝的概率模型。

图 9—1　抵押贷款申请被拒绝以及支付与收入比的散点图

他们中的 20% 将会被拒绝。

这个解释遵循了两个事实。首先,根据第 2 部分的内容,总体回归函数是给定回归因子条件下 Y 的期望值,即 $E(Y|X_1,X_2,\cdots,X_k)$。其次,根据 2.2 节内容可知,如果 Y 是一个 0～1 的二元变量,那么它的期望值(或均值)就是 $Y=1$ 时的概率,即 $E(Y)=\Pr(Y=1)$。在回归的意义下,期望值是以回归因子的值为条件的,因此估计出的概率是以 X 为条件的。所以对于一个二元变量,$E(Y|X_1,X_2,\cdots,X_k)=\Pr(Y=1|X_1,X_2,\cdots,X_k)$。简而言之,对于一个二元变量,从总体回归函数中得出的预测值,就是在给定 X 条件下 $Y=1$ 的概率。

应用于二元因变量的多元线性回归模型,又被称为线性概率模型:之所以称它是"线性的"是因为它是一条直线;之所以称它是"概率模型"是因为它建立的是一个因变量等于 1 的概率模型,在我们的例子中,即是贷款被拒绝的概率。

9.1.2　线性概率模型

线性概率模型(**linear probability model**),是第 2 部分中介绍的当因变量是二元变量而不是连续变量时的多元回归模型的称谓。因为因变量 Y 是二元变量,所以总体回归函数对应于给定 X 条件下因变量等于 1 的概率。回归因子 X 的总体系数 β_1 就是与 X 的单位变化相联系的 $Y=1$ 时的概率的变化。同理,利用所估计的回归函数计算的 OLS 预测值 $\hat{Y}_i$,就是当因变量等于 1 时的预测的概率,而 OLS 估计量 $\hat{\beta}_1$ 则估计了与 X 的单位变化相联系的当 $Y=1$ 时预测概率的变化。

第 2 部分中几乎所有的回归工具都可沿用到线性概率模型中。系数可以用 OLS 进行估计。95% 的置信区间可被构造为 ±1.96 倍的标准误;关于几个系数的假设检验可以使用第 5 章中所讨论的 F 统计量来进行;而变量之间的交互作用可以使用 6.3 节的方法来建模。由于线性概率模型的误差总是异方差的(练习 9.3),因此应该使用异方差稳健的标准误进行统计推断。

有一个不能沿用的工具是 R^2。当因变量是连续的时,设想一种 $R^2=1$ 的情形是可能的,即所有数据恰好位于直线上。当因变量是二元的时,这是不可能的,除非回归因子也是



二元的。因此，R^2 在这里并不是一个特别有用的统计量。下一节我们将会回到拟合测量指标的讨论上。

线性概率模型在重要概念9.1中总结。

在波士顿HMDA数据中的应用。用我们数据集中所有的2 380个观察值所估计的二元因变量 *deny* 对 *P/I ratio* 的OLS回归方程是：

$$\widehat{deny} = \underset{(0.032)}{-0.080} + \underset{(0.098)}{0.604}\ P/I\ ratio \tag{9.1}$$

所估计的 *P/I ratio* 的系数是正的，且总体系数在1%的水平下在统计上显著地异于0（t 统计量是6.13）。因而，债务支付占收入的比重越高的申请者其申请越有可能被拒绝。*P/I ratio* 的系数可被用来计算给定回归因子变化的条件下，拒绝概率的预测变化。例如，根据公式(9.1)，如果 *P/I ratio* 增加0.1，那么被拒绝的概率增加 $0.604 \times 0.1 \approx 0.06$，即增加6.0个百分点。

重要概念9.1

线性概率模型

线性概率模型就是下列的多元线性回归模型：

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i \tag{9.2}$$

其中，Y_i 是二元变量，因此：

$$\Pr(Y=1|X_1,X_2,\cdots,X_k) = \beta_0 + \beta_1 X_1 + \beta_2 X_2 + \cdots + \beta_k X_k$$

回归系数 β_1 就是在保持其他回归因子不变的情况下，与 X_1 的单位变化相联系的 $Y=1$ 时估计概率的变化，β_2 的含义依此类推。回归系数可以用OLS方法进行估计，并且通常的（异方差稳健的）OLS标准误可以被用来构造置信区间和假设检验。

公式(9.1)中所估计的线性概率模型可被用来计算作为 *P/I ratio* 变量函数的预测的被拒绝概率。例如，如果计划的债务支付额是申请者收入的30%，那么 *P/I ratio* 是0.3，根据公式(9.1)，预测值是 $-0.080 + 0.604 \times 0.3 = 0.101$，即根据这个线性概率模型，计划的债务支付额占收入的比重为30%的申请者，其申请具有10.1%的被拒绝概率。（这个结论不同于以图9—1中的回归线为基础得出的20%的概率，因为那条直线只用了被用于估计公式(9.1)的2 380个观测值中的127个。）

保持 *P/I ratio* 不变的条件下，种族对被拒绝概率的影响是什么？为了使问题简化，我们关注黑人申请者和白人申请者之间的差异。为了估计种族的影响，我们保持 *P/I ratio* 不变，以一个二元回归因子来扩展公式(9.1)，这个回归因子在申请者是黑人时等于1，在申请者是白人时等于0，所估计的线性概率模型是：

$$\widehat{deny} = \underset{(0.029)}{-0.091} + \underset{(0.089)}{0.559}\ P/I\ ratio + \underset{(0.025)}{0.177}\ black \tag{9.3}$$

black 的系数是0.177，表明在保持他们的 *P/I ratio* 不变的情况下，美籍非洲黑人申请者抵押贷款的申请被拒绝的概率比白人申请者高出17.7%。这个系数在1%的水平下在统计上是显著的（t 统计量是7.11）。

按字面意思，这个估计值表明在抵押贷款决策方面可能存在种族歧视，但是得出这样的结论可能为时尚早。尽管 *P/I ratio* 对信贷员的决策起作用，但是许多其他的因素也起作用，诸如申请者的潜在收入和个人信用记录。如果这些变量中的任何变量与回归因子 *black* 或 *P/I ratio* 相关，那么在公式(9.3)中忽略它们将会引起遗漏变量偏差。这样，在我们于9.3节做出

更彻底的分析之前，我们必须推迟得出任何关于抵押贷款上存在种族歧视的结论。

线性概率模型缺点。使线性概率模型易于使用的线性特征也正是它的主要缺点。再看看图 9—1：表示预测概率的那条估计线对于很低的 *P/I ratio* 值落在 0 以下，而对于较高的值则超过 1！但是，这是无意义的，因为概率不可能小于 0 或者大于 1。这个无意义的特征是线性回归不可避免的结果。为了解决这个问题，我们引入特别为二元因变量设计的新的非线性模型，即 probit 和 logit 回归模型。

9.2 probit 和 logit 回归

probit 和 logit[①] 回归是特别为二元因变量设计的非线性回归模型。因为含有二元因变量 Y 的回归建立了 $Y=1$ 时的概率模型，所以，采用迫使预测值在 0 和 1 之间的非线性表达式是有意义的。由于累积概率分布函数（c. d. f.）产生 0 和 1 之间的概率（见 2.1 节），因此它们被用在 probit 和 logit 的回归中。probit 回归使用了标准正态分布的 c. d. f.。logit 回归，也被称为 logistic 回归（logistic regression），使用了"logistic"c. d. f.。

9.2.1 probit 回归

含有单个回归因子的 probit 回归。含有单个回归因子 X 的 probit 回归模型是：

$$\Pr(Y=1|X)=\Phi(\beta_0+\beta_1 X) \tag{9.4}$$

其中，Φ 是累积标准正态分布函数（已制成表格，见附表 1）。

例如，假设 Y 是二元抵押被拒绝变量 *deny*，X 是支付额/收入额之比（*P/I ratio*），$\beta_0=-2$，$\beta_1=3$，那么如果 *P/I ratio* $=0.4$，被拒绝的概率是多少？根据公式（9.4），这个概率是 $\Phi(\beta_0+\beta_1 P/I\ ratio)=\Phi(-2+3P/I\ ratio)=\Phi(-2+3\times0.4)=\Phi(-0.8)$。根据累积正态分布表（见附表 1），$\Phi(-0.8)=\Pr(Z\leqslant-0.8)=21.2\%$。也就是说，当 *P/I ratio* 是 0.4 时，使用系数 $\beta_0=-2$ 和 $\beta_1=3$ 时 probit 模型计算出来的申请被拒绝的预测概率是 21.2%。

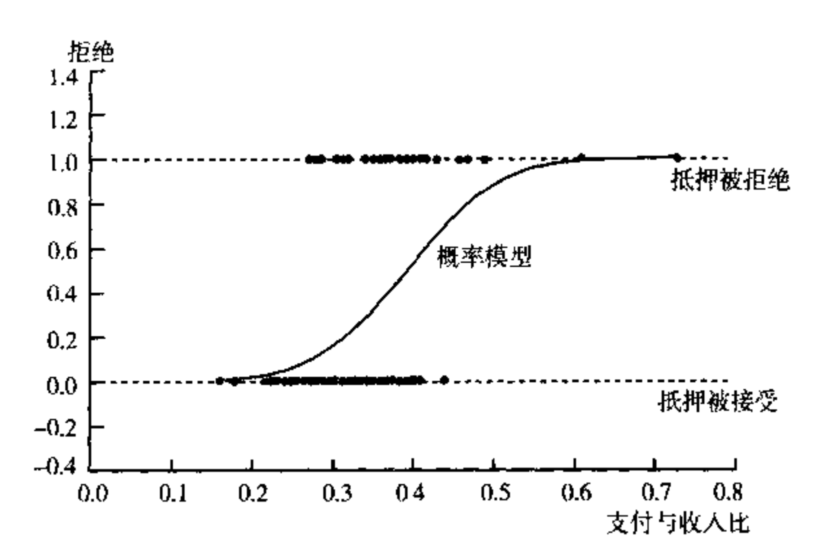
在 probit 模型中，$\beta_0+\beta_1 X$ 项在附表 1 的累积标准正态分布表中起"z"的作用。这样，前一段中的计算可以等价地按以下方式做，首先计算"z 值"，$z=\beta_0+\beta_1 X=-2+3\times0.4=-0.8$，然后查找 $z=-0.8$ 左边的正态分布尾部概率，它是 21.2%。

如果公式（9.4）中的 β_1 是正的，那么 X 增加会提高 $Y=1$ 时的概率；如果 β_1 是负的，那么 X 增加会降低 $Y=1$ 时的概率。然而，除此之外，不容易直接解释 probit 的系数 β_0 和 β_1。相反，最好通过计算概率和/或概率的变化来间接地进行解释。如果只有一个回归因子，那么解释 probit 回归最容易的方法就是用图示方法标出概率。

根据前面散点图中的 127 个观测值，图 9—2 绘制了 *deny* 对 *P/I ratio* 的 probit 回归所生成的估计回归函数。所估计的 probit 回归函数呈现伸长的"S"形：对较小的 *P/I ratio* 值而言，它接近于 0 而且是平的；对中间的值而言，它变得弯曲而且上升；而对很大值而言，它又变平了，而且接近于 1。对较小的 *P/I ratio* 值而言，被拒绝的概率很小。例如，当 *P/I ratio* $=0.2$ 时，依据图 9—2 中所估计的 probit 函数，所估计的被拒绝概率是 $\Pr(deny=1|P/I\ ratio=0.2)=2.1\%$；当 *P/I ratio* $=0.3$ 时，所估计的被拒绝概率是 16.1%；当 *P/I ratio* $=0.4$ 时，被拒绝概率陡增到 51.9%；而当 *P/I ratio* $=0.6$ 时，被拒绝概率是 98.3%。根据这个所估计的 probit 模型，对高 *P/I ratio* 值的申请者而言，被拒绝的概率接近于 1。

① 发音为 prō-bit 和 lō-jit。





注：probit 模型使用累积正态分布函数来建立给定支付与收入比（*P/I ratio*）条件下的被拒绝概率模型，或者更一般地说，建立 $\Pr(Y=1|X)$ 的模型。不像线性概率模型，probit 条件概率总是在 0 和 1 之间。

图 9—2　给定 *P/I ratio* 下，被拒绝概率的 probit 模型

多个回归因子的 probit 回归。到目前为止，我们已研究的所有回归问题都可能导致遗漏变量偏差，不过我们把一个与回归因子相关的对 Y 的决定性因素省略了，probit 回归也不例外。在线性回归中，解决的方法是把额外的变量包括进来作为回归因子，这也是解决 probit 回归中遗漏变量偏差的方法。

多个回归因子的 probit 模型，通过增加回归因子计算 z 值，将单个回归因子的 probit 模型进行了推广，因此，含有两个回归因子 X_1 和 X_2 的 probit 总体回归模型是：

$$\Pr(Y=1|X_1,X_2)=\Phi(\beta_0+\beta_1X_1+\beta_2X_2) \tag{9.5}$$

例如，假设 $\beta_0=-1.6, \beta_1=2, \beta_2=0.5$，如果 $X_1=0.4$ 且 $X_2=1$，那么 z 值是 $z=-1.6+2\times0.4+0.5\times1=-0.3$。因此，给定 $X_1=0.4, X_2=1$ 和 $Y=1$ 条件下，相应的概率就是 $\Pr(Y=1|X_1=0.4,X_2=1)=\Phi(-0.3)=38\%$。

X 变化的效应。一般地，X 变化对 Y 的效应就是 X 变化所引起的 Y 的期望变化。当 Y 是二元变量时，它的条件期望就是它等于 1 时的条件概率，因此，X 变化所引起的 Y 的期望变化就是 $Y=1$ 时的概率的变化。

回想一下 6.1 节的内容，当总体回归函数是 X 的非线性函数时，这个期望变化可分三步进行估计：第一步，利用所估计的回归函数，在 X 的初始值处计算方程的预测值；第二步，在 X 变化值 $X+\Delta X$ 处计算方程的预测值；第三步，计算这两个预测值的差。这个程序在重要概念 6.1 做了总结。正像 6.1 节中所强调的一样，不论该非线性模型如何复杂，这个方法对计算 X 变化的预测效应总是有效的。如果将重要概念 6.1 中的方法应用到 probit 模型中，那么会得出 X 变化对 $Y=1$ 时概率的估计效应。

probit 回归模型、预测概率和估计效应，在重要概念 9.2 中做了总结。

重要概念 9.2



probit 模型、预测概率和估计效应

含有多个回归因子的总体 probit 模型是：

$$\Pr(Y=1|X_1,X_2,\cdots,X_k)=\Phi(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k) \tag{9.6}$$

其中，因变量 Y 是二元变量，Φ 是累积标准正态分布函数，X_1, X_2 等是回归因子。probit 系数

β_0,β_1 等没有简单的解释方法，最好通过计算预测概率和估计回归因子变化的效应来解释这个模型。

已知 $X_1,X_2,\cdots,X_k$ 的值，$Y=1$ 时的预测概率是这样计算的：通过计算 z 值，即 $z=\beta_0+\beta_1X_1+\cdots+\beta_kX_k$，然后在正态分布表（附表1）中查找这个 z 值所对应的预测概率。

回归因子变化的估计效应是按以下步骤计算的：(1)对回归因子的初始值计算预测概率；(2)对回归因子的新值或者变化后的值计算预测概率；(3)求它们的差。

在抵押贷款数据中的应用。作为一个例子，我们拟合了数据集中2 380个观测值的抵押贷款的 *deny* 和 *P/I ratio* 的 probit 模型：

$$\widehat{\Pr(deny=1|P/I\ ratio)} = \Phi(\underset{(0.16)}{-2.19} + \underset{(0.47)}{2.97}\ P/I\ ratio) \tag{9.7}$$

估计系数 -2.19 和 2.97 的直观意义很难解释，因为它们通过 z 值影响了被拒绝概率。实际上，从公式(9.7)所估计的 probit 回归中容易断定的惟一一件事是 *P/I ratio* 与被拒绝概率正相关（*P/I ratio* 系数是正的），而且这个关系在统计上是显著的（$t=2.97/0.47=6.32$）。

但是，当 *P/I ratio* 从0.3增加到0.4时，申请被拒绝的预测概率的变化是多少？为了回答这个问题，我们遵循重要概念6.1中的方法：计算 *P/I ratio* $=0.3$ 时的被拒绝概率，然后计算 *P/I ratio* $=0.4$ 时的被拒绝概率，最后计算这两个被拒绝概率的差。当 *P/I ratio* $=0.3$ 时时，被拒绝概率是 $\Phi(-2.19+2.97\times0.3)=\Phi(-1.30)=0.097$，当 *P/I ratio* $=0.4$ 时，被拒绝概率是 $\Phi(-2.19+2.97\times0.4)=\Phi(-1.00)=0.159$，被拒绝概率的估计变化就是 $0.159-0.097=0.062$。也就是说，*P/I ratio* 从0.3增加到0.4，相应的被拒绝概率从9.7%增加到15.9%，增加了6.2个百分点。

由于 probit 回归函数是非线性的，因此 X 变化的效应依赖于 X 的初始值。例如，如果 *P/I ratio* $=0.5$，那么基于公式(9.7)所估计的被拒绝概率是 $\Phi(-2.19+2.97\times0.5)=\Phi(-0.71)=0.239$。因此，当 *P/I ratio* 从0.4增加到0.5时，预测概率的变化是 $0.239-0.159=0.080$，或8.0个百分点，大于当 *P/I ratio* 从0.3增加到0.4时增加的6.2个百分点。

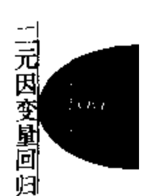
如果保持 *P/I ratio* 不变，那么种族因素对抵押贷款被拒绝概率的影响是什么？为了估计这个影响，我们估计了 *P/I ratio* 和 *black* 作为回归因子的 probit 回归方程：

$$\widehat{\Pr(deny=1|P/I\ ratio,black)} = \Phi(\underset{(0.16)}{-2.26} + \underset{(0.44)}{2.74}\ P/I\ ratio + \underset{(0.083)}{0.71}\ black) \tag{9.8}$$

虽然仍旧很难解释系数值的直观意义，但是并不难解释其符号和统计显著性。*black* 的系数是正的，表明如果保持他们的 *P/I ratio* 不变，那么美籍非洲黑人申请者的被拒绝概率比白人申请者高。这个系数在1%的水平下在统计上是显著的（*black* 系数的 t 统计量是8.55）。对 *P/I ratio* $=0.3$ 的白人申请者而言，预测的被拒绝概率是7.5%，而对 *P/I ratio* $=0.3$ 的黑人申请者而言，预测的被拒绝概率是23.3%，这两个假设申请者的被拒绝概率之差是15.8个百分点。

probit 系数的估计。这里所给出的 probit 系数是使用极大似然方法估计的。在多种应用中，包括在二元因变量回归中，极大似然方法生成有效的（最小方差）估计量。在大样本条件下，极大似然估计量是一致的且服从正态分布，因此能够用通常方法构造系数的 t 统计量和置信区间。

估计 probit 模型的回归软件典型地使用极大似然估计方法，因此，它是实际应用的一种简单方法。这样的软件所生成的标准误同样能被用做为回归系数标准误，例如，真实 probit



系数的95%的置信区间可被构造为估计系数±1.96倍的标准误。同样,使用极大似然估计量计算出的F统计量也能够用于检验联合假设。极大似然估计在9.3节中做了进一步讨论,在附录9.2中给出了另外的细节。

重要概念9.3

logit 回归

含有多个回归因子的二元因变量Y的总体logit模型是:

$$\Pr(Y=1|X_1,X_2,\cdots,X_k)=F(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)$$
$$=\frac{1}{1+e^{-(\beta_0+\beta_1X_1+\beta_2X_2+\cdots+\beta_kX_k)}} \quad (9.9)$$

除了累积分布函数不同以外,logit回归与probit回归类似。

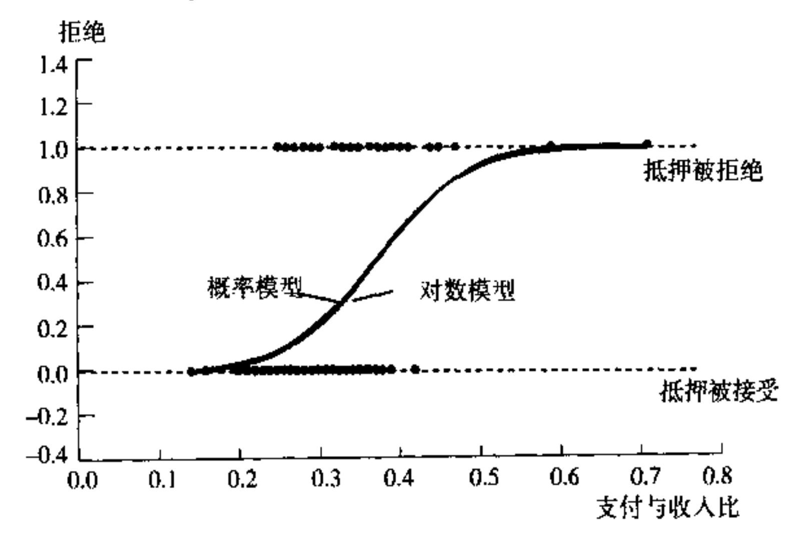
9.2.2 logit 回归

logit回归模型。logit回归模型类似于probit回归模型,区别主要是公式(9.6)中的累积标准正态分布函数Φ被累积标准logistic分布函数代替了。我们用F表示累积标准logistic分布函数,在重要概念9.3中总结了logit回归。logistic分布函数是一种用指数函数定义的特定的函数形式,如公式(9.9)中的最后一个表达式所给出的那样。

和解释probit系数一样,最好通过计算预测的概率以及预测的概率之差来解释logit系数。

logit模型的系数可以用极大似然法估计。在大样本条件下,极大似然估计量是一致的且正态分布的,因此可以用通常方法构造系数的t统计量和置信区间。

logit回归函数与probit回归函数是相似的,图9—3证明了这一点。图9—3描绘出了使用与图9—1和图9—2中相同的127个观察值,利用极大似然法所估计的因变量*deny*和单个回归因子*P/I ratio*的probit与logit回归函数。这两个函数之间的差异很小。



注:给定支付与收入比(*P/I ratio*)条件下,这些probit和logit模型生成了几乎相同的抵押贷款申请被拒绝的概率估计值。

图9—3 给定*P/I ratio*条件下,被拒绝概率的probit和logit模型

历史上,logit回归发展的主要动因是logistic累积分布函数可以比累积正态分布函数计



9.3.1 非线性最小二乘估计

像 probit 模型的系数一样，当未知系数以非线性方式进入总体回归函数时，非线性最小二乘是估计回归函数未知系数的一般方法。

让我们回顾 5.3 节中关于多元线性回归模型系数的 OLS 估计量的讨论。OLS 估计量使公式(5.8)中的预测误差平方和 $\sum_{i=1}^{n}[Y_i-(b_0+b_1X_{1i}+\cdots+b_kX_{ki})]^2$ 达到最小。原则上说，可以通过检查 $b_0,b_1,\cdots,b_k$ 的多次试验值并确定那些能够使误差平方和最小的值来计算 OLS 估计量。

可以使用同样的方法估计 probit 系数。由于该回归模型是关于系数的非线性关系，因此这种方法被称为非线性最小二乘。对一组试验系数值 b_0,b_1 等，构造预测误差平方和：

$$\sum_{i=1}^{n}[Y_i-\Phi(b_0+b_1X_{1i}+\cdots+b_kX_{ki})]^2 \tag{9.11}$$

上式与线性回归模型的误差平方和是一样的，只是这里的回归函数是由 probit 模型给出的。probit 模型系数的**非线性最小二乘估计量**(nonlinear least squares estimator)，就是使表达式(9.11)中的预测误差平方和最小的 $b_0,b_1,\cdots,b_k$ 的值。同理，以 logistic 分布函数 F 代替表达式(9.11)中的 Φ，logit 模型系数的非线性最小二乘估计量可使预测误差平方和达到最小。

在线性回归模型中，很幸运有一个公式将 OLS 估计量表示成数据的函数。不幸的是，对 probit 模型却没有这样的公式，因此非线性最小二乘估计量的值必须由计算机来求。回归分析的软件已经把求解这个最小化问题的复杂算法的计算程序置入其中，因此，在实践中计算非线性最小二乘估计量的工作可大大减少。

probit 系数的非线性最小二乘估计量拥有线性回归中 OLS 估计量的两个重要性质：它是一致的(随着样本容量的增大，它逼近真值的概率接近 1)且在大样本条件下是正态分布的。不过，还存在比非线性最小二乘估计量有更小方差的估计量，也就是说，非线性最小二乘估计量是无效的估计量。由于这个原因，probit 系数的非线性最小二乘估计量在实际中很少被使用，而常用极大似然法来估计参数。

9.3.2 极大似然估计

似然函数(likelihood function)是数据的联合概率分布，可将其看做是未知系数的函数。未知系数的**极大似然估计量**(maximum likelihood estimator，简写为 MLE)是指那些使似然函数最大化的系数值。由于 MLE 是选择那些使似然函数最大化的未知系数，而似然函数反过来又是该数据的联合概率分布，因此实际上，MLE 选择的参数的值，是使实际被观测到的数据被抽取的概率最大化的参数值。在这个意义下，MLE 就是“最可能”生成该数据的参数值。

为了举例说明极大似然估计，考虑一个没有回归因子的二元因变量的两个独立同分布的观测值 Y_1 和 Y_2。这里，Y 是一个贝努里(Bernoulli)随机变量，而惟一要估计的未知参数是 $Y=1$ 的概率 p，它也是 Y 的均值。

为了得到极大似然估计量，我们需要似然函数的表达式，这反过来又需要数据的联合概率分布表达式。上面的两个观察值 Y_1 和 Y_2 的联合概率分布是 $\Pr(Y_1=y_1,Y_2=y_2)$。因为 Y_1 和 Y_2 是独立分布的，联合分布是两个个别分布的积(公式(2.21))，所以 $\Pr(Y_1=y_1,Y_2=$

$y_2)=\Pr(Y_1=y_1)\Pr(Y_2=y_2)$。贝努里分布可以用下式来归纳：$\Pr(Y=y)=p^y(1-p)^{1-y}$，其中当 $y=1$ 时，$\Pr(Y=1)=p^1(1-p)^0=p$；当 $y=0$ 时，$\Pr(Y=0)=p^0(1-p)^1=1-p$。这样，Y_1 和 Y_2 的联合概率分布就是 $\Pr(Y_1=y_1,Y_2=y_2)=[p^{y_1}(1-p)^{1-y_1}]\times[p^{y_2}(1-p)^{1-y_2}]=p^{(y_1+y_2)}(1-p)^{2-(y_1+y_2)}$。

似然函数就是该联合概率的分布，可把它看做是未知系数的函数。对贝努里随机变量的 $n=2$ 个独立同分布观测值而言，似然函数是：

$$f(p;Y_1,Y_2)=p^{(Y_1+Y_2)}(1-p)^{2-(Y_1+Y_2)} \tag{9.12}$$

p 的极大似然估计量就是使公式(9.12)中的似然函数最大化的 p 值。同所有的最大化和最小化问题一样，这个问题可通过反复试验来解决，也就是说，你可以试不同的 p 值并计算相应的似然函数 $f(p;Y_1,Y_2)$，直至你相信你已使这个函数最大化为止。不过，在这个例子中，使用微积分来使似然函数最大化会生成一个 MLE 的简单公式：MLE 就是 $\hat{p}=\frac{1}{2}(Y_1+Y_2)$。换句话说，$p$ 的 MLE 正好是样本平均值！事实上，对于一般的 n，贝努里概率 p 的 MLE 的 $\hat{p}$ 就是样本均值，即 $\hat{p}=\overline{Y}$(这在附录 9.2 中证明)。在这个例子中，MLE 就是通常的 p 的估计量，即样本中 $Y_i=1$ 时次数的比率。

这个例子与估计 probit 和 logit 回归模型未知系数问题很类似。在那些模型中，成功概率 p 并不是常数，而是依赖于 X，也就是说，它是以 X 为条件的成功概率，它在 probit 模型方程(9.6)中和 logit 模型方程(9.9)中已给出。因此，probit 和 logit 似然函数与公式(9.12)中的似然函数相似，除了其中的成功概率随观测值的变化而变化(因为它依赖于 X_i)之外。附录 9.2 中给出了 probit 和 logit 似然函数的表达式。

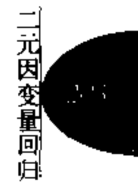
像非线性最小二乘估计一样，在大样本条件下，MLE 也是一致性的和正态分布的。因为回归软件通常计算 probit 系数的 MLE，所以实践中很容易使用这个估计量。本章所给出的所有 probit 和 logit 模型的估计系数都是 MLE。

基于 MLE 的统计推断。由于在大样本条件下，MLE 是正态分布的，因此基于 MLE 的 probit 和 logit 系数的统计推断，与基于 OLS 估计量的线性回归函数系数的推断方法相同，也就是说，可使用 t 统计量进行假设检验，可使用 95% 的置信水平来构造 ±1.96 倍标准误的置信区间。可使用 F 统计量对多个系数的联合假设进行检验，这在某种程度上与第 5 章所讨论的线性回归模型的检验类似。所有这些都和线性回归模型中的统计推断完全类似。

在实践中的一个重要问题是，某些统计软件使用 F 统计量报告联合假设的检验结果，而另一些软件则使用卡方统计量报告检验结果。卡方统计量是 $q\times F$，其中 q 是被检验的约束条件的个数。因为在零假设下，在大样本中 F 统计量服从 χ_q^2/q 分布，所以 $q\times F$ 服从 χ_q^2 分布。由于这两个方法的区别只是是否被 q 来除，因此它们得到的推断是相同的，但是你需要知道软件中执行的是哪个方法，以便你使用正确的临界值。

9.3.3 拟合优度测量

在 9.1 节中，我们曾经提到 R^2 对于线性概率模型是个不好的拟合测量。这个结论也同样适用于 probit 和 logit 回归。对于二元因变量模型，有两个拟合测量，即"被正确地预测的比重"和"伪 R^2"。被正确地预测的比重(fraction correctly predicted)使用如下规则：如果 $Y_i=1$ 且预测概率超过50%，或者如果 $Y_i=0$ 且预测概率小于 50%，那么就说 Y_i 被正确地预测了；否则就说 Y_i 被不正确地预测了。"被正确地预测的比重"就是在 n 个观测值 $Y_1,\cdots,Y_n$ 中被正确地预测的部分所占的比重。



这个拟合测量的优点是易于理解,缺点是它没有反映预测的质量:如果 $Y_i=1$,那么无论预测概率是 51% 还是 90%,观测值都被看做是被正确地预测了。

伪 R^2(pseudo $-R^2$)利用似然函数测量了模型的拟合程度。由于 MLE 最大化了似然函数,因此,就像在 OLS 估计的线性回归中增加回归因子一定会减少残差平方和一样,向 probit 和 logit 模型中添加额外的回归因子也一定会增大最大化似然函数的值。这表明,通过比较含有全部回归因子的最大化似然函数的值和无回归因子的似然函数的值,可以测量 probit 模型的拟合质量,实际上这就是伪 R^2 所做的工作。附录 9.2 中给出了伪 R^2 的计算公式。

9.4 在波士顿 HMDA 数据案例中的应用

前两节的回归分析表明,黑人申请者的被拒绝率要比白人申请者的被拒绝率高,如果保持他们的 *P/I ratio* 不变。不过,当信贷员决定一项抵押贷款申请时,他们会合理地权衡许多因素,如果那些其他因素中的任何一个因素因种族因素而系统地不同,那么到目前为止所考虑的估计量就会有遗漏变量偏差。

在本节中,我们更进一步地了解一下在波士顿 HMDA 数据中是否存在种族歧视的统计证据。具体来说,我们的目的就是要在保持当一个信贷员决定一项抵押贷款申请时所合理考虑的申请者的特征不变的请况下,估计种族因素对被拒绝概率的影响。

在波士顿 HMDA 数据集中,信贷员通过抵押贷款申请可以获得的最重要的变量列在表 9—1 中,这些变量都是我们在贷款决策实证模型的分析中将集中关注的变量。前两个变量是对所提出的一笔贷款对贷款申请者造成的财务负担的直接测量,以他(或她)的收入来测量。第一个变量是 *P/I ratio*;第二个变量是与购房相关的费用和收入之比。接下来的变量是贷款额与房屋评估价值的比值。如果贷款与价值的比接近 1,那么如果申请者拖欠贷款而且银行取消了抵押品的赎回权,银行在获得全部贷款额补偿方面将遇到麻烦。最后三个财务变量总结了申请者的信用史。如果一个申请者过去在偿清债务方面一直不很可靠,那么该信贷员会合理地考虑该申请者将来是否有能力支付该笔抵押贷款,或者他或她是否想在将来归还该笔贷款。这三个变量测量了不同类型的信用史,信贷员在对这三个变量进行分析时可能分别给予不同的权重。第一个变量涉及消费者信用,比如信用卡债务;第二个变量涉及以前的抵押支付史;第三个变量测量了那些非常严重的信用问题,这些信用问题已经在公开的法律档案做了记载,例如呈请破产备案。

表 9—1 还列出了对信贷员决策起重要作用的其他几个变量。有时申请者必须申请私人抵押保险。[①] 信贷员需要知道该保险申请者是否被拒绝,以及这种拒绝是否会对信贷员的决策产生负面影响。接下来的三个变量与申请者的预期偿还能力有关,它们涉及申请者的就业状况、婚姻状况和受教育水平。一旦取消抵押品赎回权,那么财产特征也是很重要的,下一个变量表明了财产是否是共管的。表 9—1 中的最后两个变量一个是申请者是黑人还是白人,另一个是申请是否被拒绝。在这些数据中,14.2% 的申请者是黑人,12.0% 的申请被拒绝。

表 9—2 给出了基于这些变量的回归结果。第(1)~第(3)列中所给出的基准设定包括表 9—1 中的财务变量、表示私人抵押保险是否被拒绝的变量以及申请者是否是自我就业的

① 抵押保险是这样一种保险政策,即如果借款人违约,那么保险公司将为银行制作月度支付保险单。在本案例的研究期间,如果贷款与价值比超过 80%,那么一般要求申请者购买抵押保险。

表 9—1　　**抵押贷款决策的回归模型所包括的变量**

变量名称	定义	样本均值
财务变量		
P/I ratio	月度总债务支付额与月度总收入之比	0.331
Housing expense-to-income ratio	月度住房费用与月度总收入之比	0.255
Loan-to-value ratio	贷款规模与财产评估价值之比	0.738
Consumer credit score	1:如果没有"延迟"支付或者不良行为 2:如果存在一次或两次延迟支付或不良行为 3:如果有超过两次的延迟支付行为 4:如果对于决策存在不足的信用史 5:如果存在超过 60 天的支付不良行为信用史 6:如果存在超过 90 天的支付不良行为信用史	2.1
Mortgage credit score	1:如果没有延迟的抵押支付行为 2:如果没有抵押支付史记录 3:如果有一次或两次延迟的抵押支付行为 4:如果存在超过两次的延迟支付行为	1.7
Public bad credit record	如果存在任何信用问题的公开记录(破产、亏损、代管行为等),变量值为 1;否则为 0	0.074
其他的申请者特征		
Denied mortgage insurance	如果申请者曾申请抵押保险且被拒绝,变量值为 1;否则为 0	0.020
Self-employed	如果自我就业,变量值为 1;否则为 0	0.116
Single	如果申请者是单身,变量值为 1;否则为 0	0.393
High school diploma	如果申请者高中毕业,变量值为 1;否则为 0	0.984
Unemployment rate	1989 年马萨诸塞州申请者所在行业中的失业率	3.8
Condominium	如果单位是共管的,变量值为 1;否则为 0	0.288
Black	如果申请者是黑人,变量值为 1;如果申请者是白人则为 0	0.142
Deny	如果抵押贷款申请被拒绝,变量值为 1;否则为 0	0.120

变量。信贷员通常使用贷款与价值比的门限值或切割值,所以该变量的基准设定使用贷款与价值比是高的(≥0.95)、中等的(在 0.8 和 0.95 之间)还是低的(<0.8;为了避免完全多重共线性,这种情况被省略了)二元变量。前三列中的回归因子与波士顿联邦储备银行在它们最初对这些数据的分析中所考虑的基准设定中的回归因子相似。[①] 第(1)~第(3)列中的回归只在如何建立被拒绝概率的模型上有所不同,它们分别使用的是线性概率模型、logit 模型和 probit 模型。

① 第(1)~第(3)列中的回归因子与 Munnell 等(1996)文中表2(1)的回归因子之间的差异是:Munnell 等文中的回归因子包括了家庭位置、贷款人身份以及不能公开得到的数据的指标;还有一个反映房屋是否为多家共用的指标,这个指标在这里并不重要,因为我们主要关注单个家庭的住房情况;还有一个指标是反映净财富的变量,不过这里我们省略了,因为这个变量有一些很大的正值和负值,所以根据这些数据做出的结论将对这些特别的"异常"观测值非常敏感,这是有风险的。



表 9—2　　　　使用波士顿 HMDA 数据的抵押贷款被拒绝的回归模型

因变量:如果抵押贷款申请被拒绝,那么 *deny* =1;否则,*deny* =0;2 380 个观测值

回归模型	LPM	Logit	Probit	Probit	Probit	Probit
回归因子	(1)	(2)	(3)	(4)	(5)	(6)
Black	0.084**	0.688**	0.389**	0.371**	0.363**	0.246
	(0.023)	(0.182)	(0.098)	(0.099)	(0.100)	(0.448)
P/I ratio	0.449**	4.76**	2.44**	2.46**	2.62**	2.57**
	(0.114)	(1.33)	(0.61)	(0.60)	(0.61)	(0.66)
Housing expense-to-income ratio	-0.048	-0.11	-0.18	-0.30	-0.50	-0.54
	(0.110)	(1.29)	(0.68)	(0.68)	(0.70)	(0.74)
Medium loan-to-value ratio	0.031*	0.46**	0.21**	0.22**	0.22**	0.22**
(0.80≤*loan-value ratio*≤0.95)	(0.013)	(0.16)	(0.08)	(0.08)	(0.08)	(0.08)
High loan-to-value ratio	0.189**	1.49**	0.79**	0.79**	0.84**	0.79**
(*loan-value ratio*≥0.95)	(0.050)	(0.32)	(0.18)	(0.18)	(0.18)	(0.18)
Consumer credit score	0.031**	0.29**	0.15**	0.16**	0.34**	0.16**
	(0.005)	(0.04)	(0.02)	(0.02)	(0.11)	(0.02)
Mortgage credit score	0.021	0.28*	0.15*	0.11	0.16	0.11
	(0.011)	(0.14)	(0.07)	(0.08)	(0.10)	(0.08)
Public bad credit record	0.197**	1.23**	0.70**	0.70**	0.72**	0.70**
	(0.035)	(0.20)	(0.12)	(0.12)	(0.12)	(0.12)
Denied mortgage insurance	0.702**	4.55**	2.56**	2.59**	2.59**	2.59**
	(0.045)	(0.57)	(0.30)	(0.29)	(0.30)	(0.29)
Self-employed	0.060**	0.67**	0.36**	0.35**	0.34**	0.35**
	(0.021)	(0.21)	(0.11)	(0.11)	(0.11)	(0.11)
Single				0.23**	0.23**	0.23**
				(0.08)	(0.08)	(0.08)
High school diploma				-0.61**	-0.60**	-0.62**
				(0.23)	(0.24)	(0.23)
Unemployment rate				0.03	0.03	0.03
				(0.02)	(0.02)	(0.02)
Condominium					-0.05	
					(0.09)	
Black × P/I ratio						-0.58
						(1.47)
Black × housing expense-to-income						1.23
ratio						(1.69)
其他的信用等级指示变量	否	否	否	否	是	否
Constant	-0.183**	-5.71**	-3.04**	-2.57**	-2.90**	-2.54**
	(0.028)	(0.48)	(0.23)	(0.34)	(0.39)	(0.35)
检验排除变量组的 *F* 统计量和 *P* 值	(1)	(2)	(3)	(4)	(5)	(6)
Applicant single; HS diploma;				5.85	5.22	5.79
industry unemployment rate				(<0.001)	(0.001)	(<0.001)
Additional credit rating					1.22	
indicator variables					(0.291)	
Race interactions and black						4.96
						(0.002)
Race interaction only						0.27
						(0.766)
白人申请者和黑人申请者被拒绝的预测概率之差(百分点)	8.4%	6.0%	7.1%	6.6%	6.3%	6.5%

注:这些回归是使用附录 9.1 中所描述的波士顿 HMDA 数据集中的 n = 2 380 个观测值估计的。线性概率模型用 OLS 估计,而 probit 和 logit 回归用极大似然法估计。标准误在系数下面的括号中给出,p 值在 F 统计量下面的括号中给出。最后一行中的预测概率的变化,是针对除种族因素外回归因子的值都等于样本均值的这样一个假想申请者而计算的。个别系数在*5%或**1%的显著性水平下在统计上是显著的。

上是联合地显著的，因此种族仍旧有显著的影响。此外，所估计的被拒绝概率的种族差异(6.5%)实质上与其他 probit 回归中的结果是相同的。

在所有 6 个设定中，如果保持申请者的其他特征不变，那么种族因素对被拒绝概率的影响在 1% 的水平下在统计上是显著的。黑人申请者和白人申请者之间被拒绝概率的估计差异在 6.0 个百分点到 8.4 个百分点之间变动。

了解评价这个差异是大还是小的方法，需要回到本章开头所提出的那个问题的一个变异上。假设有两个抵押贷款的申请个人：一个白人和一个黑人，但是在回归(3)中的其他自变量却有相同的值。具体来说，除了种族因素之外，回归(3)中其他自变量的值都是 HMDA 数据集合中的样本平均值。白人申请者面临着 7.4% 的被拒绝的可能性，但黑人申请者则面临着 14.5% 的被拒绝的可能性。所估计的被拒绝概率的种族差异为 7.1 个百分点，这意味着黑人申请者被拒绝的可能性几乎是白人申请者的 2 倍。

表 9—2 中(以及波士顿联邦调查署最初的研究)的结果，提供了抵押贷款被拒绝在种族差异上的统计证据，按照法律这是不应该存在的。这个证据在督促银行监管者做出政策变化方面起了重要作用，[①]但是经济学家喜欢有益的争论，这些结论也会引发激烈的争论，这不足为奇。

由于在借贷方面存在种族歧视的提议被指控，因此我们简短地回顾一下这个争论的一些论点。为此，采用第 7 章提出的框架结构是很有用的，也就是说，考虑表 9—2 中结果的内部有效性和外部有效性，这是前面波士顿 HMDA 数据分析的代表。由许多对波士顿联邦储备银行最初的研究所构成的批判涉及内部有效性：数据中可能的误差，不同的非线性函数形式，额外的交互作用等等。通过对原始数据的仔细审察，发现了一些错误，而这里所给出的结果(和波士顿联邦调查署最终公布的研究)都是以“清洁过的”数据集为基础的。其他设定的估计——不同函数形式和/或额外的回归因子——也得到了与表 9—2 中的那些结果可比较的种族差异的估计值。关于内部有效性的一个潜在的更难的问题是，在个人贷款面对面会谈期间是否能够得到重要的非种族方面的财务信息，而这些信息没有在贷款申请过程中被记录下来，而又与种族因素相关，如果是这样的话，在表 9—2 的回归中仍然有可能存在遗漏变量偏差。最后，有人对外部有效性提出了质疑：即使 1990 年在波士顿存在种族歧视，由此暗含着当今其他地方的贷款人中存在种族歧视也是错误的。解决外部有效性问题的惟一方法就是考虑取自其他地方和年份的数据。[②]

9.5 结论

当因变量是二元变量时，总体回归函数就是以回归因子为条件当 $Y=1$ 时的概率。对这个总体回归函数的估计需要找到对其概率解释做出公正评价的函数形式，估计那个函数的未知参数，并解释结果，所得出的预测值就是预测概率，某个回归因子 X 的变化的估计效应就是由 X 的变化所引起的当 $Y=1$ 时的概率变化。

给定回归因子，建立 $Y=1$ 的概率模型的一个很自然的方法是使用累积分布函数，其中，累积分布函数的变量依赖于回归因子。probit 回归使用正态累积分布函数作为回归函数，



① 这些政策变化包括由联邦银行监管者制定的公平借贷审核方式的变化、美国司法部所做出的在调查询问上的变化，以及银行与其他住房贷款发起公司的教育水平提高计划。

② 如果你对进一步地阅读这个主题感兴趣，那么一个很好的出发点是 1998 年春季的《经济展望杂志》(*Journal of Economic Perspectives*)中关于种族歧视和经济学的专题论丛。在那个专题论丛中，由 Helen Ladd(1998)所做的论文研究了关于抵押贷款中种族歧视的证据和争论。Goering 与 Wienk(1996)给出了更详细的讨论。

而 logit 回归使用 logistic 累积分布函数作为回归函数。由于这些模型是未知参数的非线性函数,因此估计这些参数要比估计线性回归系数复杂得多。标准的估计方法是极大似然法。在实际中,使用极大似然估计的统计推断与多元线性回归中统计推断的方法相同。例如,系数的 95% 的置信区间被构造为估计系数的 ±1.96 倍的标准误。

一般兴趣框

诺贝尔经济学奖获得者 James J. Heckman 和 Daniel L. McFadden

2000 年度诺贝尔经济学奖被同时授予了两位经济计量学家,芝加哥大学的 James J. Heckman 和加利福尼亚大学柏克莱分校的 Daniel L. McFadden,因为他们对个人和企业数据分析所做出的基础性贡献。他们的大部分工作解决了因受限因变量而产生的困难。

Heckman 因为发展了处理样本选择的工具而被授予诺贝尔经济学奖。如同 7.2 节中所讨论的,当数据的可获得性受到与因变量的值有关的选择过程影响时,样本选择偏差就会出现。例如,假设你想使用一个取自总体的随机样本来估计收入和一些回归因子 X 之间的关系。如果你使用已就业工人——即那些报告了正收入的工人——的子样本估计回归,那么 OLS 估计值可能受选择偏差影响。Heckman 的解决方法就是设定一个含有指明工人是否在劳动力队伍中(是否在子样本中)的二元因变量的初始方程,并把这个方程和收入方程看做是一个联立方程组。这个一般性的策略已被推广到许多领域中出现的选择问题上,其范围从劳动经济学,到产业组织学,再到金融学。

McFadden 因为发展了分析离散选择数据的模型(一个高中毕业生是参军、上大学还是工作?)而被授予这个奖。他是从一个人使其每种可能选择的期望效用最大化这一问题出发进行考虑的,每种可能选择依赖于可观测的变量(如工资、工作特征和家庭背景)。然后,他导出了含有未知系数的个体选择概率模型,反过来这些模型可以使用极大似然法进行估计。在许多领域中,包括劳动经济学、卫生经济学以及运输经济学领域,这些模型及其推广已被证明在分析离散选择数据上是非常有用的。

要想获得关于他们或者其他诺贝尔经济学奖获得者的更多信息,请访问诺贝尔基金会网站:www.nobel.se/economics。

尽管存在内在的非线性特征,但有时总体回归函数能被线性概率模型即多元线性回归所生成的直线充分地逼近。线性概率模型、probit 回归模型以及 logit 回归模型,被应用于波士顿 HMDA 数据中时,都给出了相似的"底线"答案:所有的三个方法都估计出了(在其他条件都类似的情况下)黑人申请者和白人申请者在抵押贷款申请被拒绝率上的实质性差异。

二元因变量是最普通的受限因变量的例子,受限因变量是指具有有限范围的因变量。20 世纪最后的 25 年,在分析其他受限因变量的经济计量方法方面取得了重要的进展(见本章一般兴趣框),附录 9.3 中回顾了当中的一些方法。

总结



1. 当 Y 是二元变量时,多元线性回归模型被称为线性概率模型。总体回归线表示给定回归因子 $X_1, X_2\cdots, X_k$ 的值的条件下 $Y=1$ 的概率。

2. probit 模型和 logit 模型都是当 Y 是二元变量时所使用的非线性回归模型。不像线性概率模型,probit 回归和 logit 回归可以确保对于所有的 X 值当 $Y=1$ 时其预测概率在 0 和 1

之间。

3. probit 回归使用标准正态累积分布函数。logit 回归使用 logistic 累积分布函数。logit 和 probit 的系数使用极大似然法进行估计。

4. probit 回归和 logit 回归中的系数值不容易解释。与一个或者多个 X 变量的变化相联系的 $Y=1$ 的概率的变化,可以使用重要概念 6.1 中列出的非线性模型的一般方法进行计算。

5. 在线性概率模型、logit 模型以及 probit 模型中,系数的假设检验可以使用通常的 t 统计量和 F 统计量来进行。

重要术语

受限因变量　线性概率模型　probit(概率单位模型)　logit(对数单位模型)　非线性最小二乘估计量　似然函数　极大似然估计量(MLE)　被正确地预测的比率　伪 R^2

复习概念

9.1 假设一个线性概率模型生成一个等于 1.3 的 Y 的预测值,请说明为什么这个值是没有意义的。

9.2 在表 9—2 中,第(1)列中关于 *black* 的估计系数是 0.084,第(2)列中是 0.688,第(3)列中是 0.389。尽管有这么大的差别,但是 3 个模型都得出了种族对抵押贷款被拒绝概率的边际效应的相似估计值,为什么会是这样呢?

9.3 你的一个朋友正考虑使用个人数据来研究你们学校吸烟的决定性因素。她问你,她是该使用 probit 模型、logit 模型还是线性概率模型。你会给她什么建议?为什么?

9.4 为什么使用极大似然法而不是 OLS 法来估计 probit 模型和 logit 模型的系数?

练习

*9.1 使用公式(9.8)中所估计的 probit 模型回答下列问题:

a. 一个黑人抵押贷款申请者的 *P/I ratio* 值是 0.35,他的申请被拒绝的概率是多少?

b. 假设该申请者将这个比率减小到 0.30,这会对他的抵押贷款被拒绝的概率有什么影响?

c. 如果该申请者是一个白人,请重新回答(a)和(b)中的问题。

d. *P/I ratio* 对抵押贷款被拒绝概率的边际效应是否依赖于种族因素?请解释说明。

9.2 使用公式(9.10)中的 logit 模型重新回答问题 9.1。logit 和 probit 的结果是否相似?请解释说明。

9.3 考虑下列线性概率模型:$Y_i=\beta_0+\beta_1X_i+u_i$,其中 $\Pr(Y_i=1|X_i)=\beta_0+\beta_1X_i$。

a. 证明:$E(u_i|X_i)=0$。



b. 证明:$\mathrm{var}(u_i|X_i)=(\beta_0+\beta_1X_i)[1-(\beta_0+\beta_1X_i)]$(提示:复习公式(2.7))

c. u_i 是异方差的吗?请解释说明。

d. (复习 9.3 节)推导似然函数。

9.4 利用表 9—2 中第(1)列所估计的线性概率模型回答下列问题:

a. 有两个申请者:一个黑人和一个白人,同时申请一项抵押贷款。除了种族因素之外,对其他所有回归因子的值都相同。对于黑人申请者来说,他的抵押贷款的申请被拒绝的可能性比白人申请者会大多少?

b. 就(a)中的答案,构造95%的置信区间。

c. 考虑一个可能会使(a)中的答案产生偏差的重要遗漏变量。它会是什么呢?它是如何使结果产生偏差的?

9.5 (需要9.3节的内容和微积分的知识)假设一个随机变量 Y 具有如下的概率分布:$\Pr(Y=1)=p$, $\Pr(Y=2)=q$,而 $\Pr(Y=3)=1-p-q$。从这个分布中抽取容量为 n 的一个随机样本,随机变量用 $Y_1,Y_2,\cdots,Y_n$ 表示。

a. 推导参数 p 和 q 的似然函数。

b. 推导 p 和 q 的 MLE 公式。

附录 9.1 波士顿 HMDA 数据集

波士顿 HMDA 数据集是由波士顿联邦储备银行的研究人员搜集的。该数据集合并了抵押贷款申请的信息以及收到这些抵押贷款申请的银行与其他贷款机构的追踪调查信息。该数据涉及波士顿大区在1990年的抵押贷款申请。完全的数据集共有2 925个观测值,由黑人和西班牙人的所有抵押贷款申请加上一个白人抵押贷款申请的随机样本所构成。

为了缩小本章分析的范围,我们使用只有一个住房的家庭(因此,排除了多个住房的家庭的数据),并且只针对黑人申请者和白人申请者(这样,也排除了其他少数民族群体的数据)的数据的子集,这样就只剩下了2 380个观测值。表9—1给出了本章中所使用的变量的定义。

很感谢波士顿联邦储备银行研究部的 Geoffrey Tootell 向我们提供了这些数据。关于这个数据集的更多信息,连同波士顿联邦储备银行研究人员所得出的结论一起,在 Alicia H. Munnell, Geoffrey M. B. Tootell, Lynne E. Browne 与 James McEneaney 等的论文"Mortgage Lending in Boston: Interpreting HMDA Data"(*American Economic Review*, 1996, pp. 25-53)中可找到。

附录 9.2 极大似然估计

本附录提供了与本章中所讨论的二元响应模型的内容有关的极大似然估计的简要介绍。我们从推导 n 个独立同分布的贝努里随机变量观察值的成功概率 p 的 MLE 开始。然后,我们转向 probit 模型和 logit 模型,并讨论伪 R^2。我们以预测概率的标准误的讨论作为结束。本附录使用两点积分。

n 个 i. i. d. 贝努里随机变量的 MLE



计算 MLE 的第一步是推导联合概率分布。对 n 个 i. i. d. 的贝努里随机变量观测值而言,这个联合概率分布是9.3节中 $n=2$ 的情形到一般的 n 的扩展:

$$\begin{aligned}\Pr(Y_1&=y_1,Y_2=y_2,\cdots,Y_n=y_n)\\&=[p^{y_1}(1-p)^{(1-y_1)}]\times[p^{y_2}(1-p)^{(1-y_2)}]\times\cdots\times[p^{y_n}(1-p)^{(1-y_n)}]\\&=p^{y_1+\cdots+y_n}(1-p)^{n-(y_1+\cdots+y_n)}\end{aligned}\tag{9.13}$$

似然函数就是该联合概率分布，被看做未知系数的函数。假设 $S=\sum_{i=1}^{n}Y_i$，那么似然函数就是：

$$f_{Bernoulli}(p;Y_1,Y_2,\cdots,Y_n)=p^S(1-p)^{n-S} \tag{9.14}$$

p 的 MLE 就是使公式(9.14)中的似然函数最大的 p 值，可以使用微积分方法求似然函数的最大值，求似然函数对数的最大值，而不直接求似然函数的最大值，这要简便得多(因为对数是严格的增函数，所以使似然函数或者它的对数最大化会得出相同的估计量)。对数似然函数就是 $S\ln(p)+(n-S)\ln(1-p)$，而且对数似然函数关于 p 的导数是：

$$\frac{d}{dp}\ln[f_{Bernoulli}(p;Y_1,Y_2,\cdots,Y_n)]=\frac{S}{P}-\frac{n-S}{1-p} \tag{9.15}$$

设公式(9.15)中的导数为 0，求解 p 得出 MLE$\hat{p}=S/n=\overline{Y}$。

probit 模型的 MLE

对于 probit 模型，以 $X_{1i},X_{2i},\cdots,X_{ki}$ 为条件，$Y=1$ 时的概率是 $p_i=\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})$。第 i 个观测值的条件概率分布就是 $\Pr(Y_i=y_i|X_{1i},\cdots,X_{ki})=p_i^{y_i}(1-p_i)^{1-y_i}$。假设$(X_{1i},\cdots,X_{ki},Y_i)$是 i. i. d. 的，$i=1,2,\cdots,n$，那么以变量 X 为条件，$Y_1,\cdots,Y_n$ 的联合概率分布是：

$$\begin{aligned}&\Pr(Y_1=y_1,\cdots,Y_n=y_n|X_{1i},\cdots,X_{ki},i=1,\cdots,n)\\&=\Pr(Y_1=y_1|X_{11},\cdots,X_{k1})\times\cdots\times\Pr(Y_n=y_n|X_{1n},\cdots,X_{kn})\\&=p_1^{y_1}(1-p_1)^{1-y_1}\times\cdots\times p_n^{y_n}(1-p_n)^{1-y_n}\end{aligned} \tag{9.16}$$

似然函数就是该联合概率分布，被看做未知系数的函数。考虑似然函数的对数会简便一些，因此，对数似然函数是：

$$\begin{aligned}&\ln[f_{probit}(\beta_0,\cdots,\beta_k;Y_1,\cdots,Y_n|X_{1i},\cdots,X_{ki},i=1,\cdots,n)]\\&=\sum_{i=1}^{n}Y_i\ln[\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})]+\\&\quad\sum_{i=1}^{n}(1-Y_i)\ln[1-\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})]\end{aligned} \tag{9.17}$$

其中，该表达式涵盖了 probit 的条件概率公式：$p_i=\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})$。

probit 模型的 MLE 使似然函数最大化，或者说，使公式(9.17)中所给出的对数似然函数最大化。由于没有简单公式计算 MLE，因此 probit 似然函数必须使用数值算法在计算机上实现最大化。

在一般条件下，极大似然估计量是一致的，而且在大样本条件下有正态的抽样分布。

logit 模型的 MLE

logit 模型的似然函数可按照与 probit 模型的似然函数相同的方法进行推导，惟一的差别就是 logit 模型的条件成功概率 p_i 由公式(9.9)给出。因此，logit 模型的对数似然函数由公式(9.17)给出，用$[1+e^{-(\beta_0+\beta_1X_{1i}+\beta_2X_{2i}+\cdots+\beta_kX_{ki})}]^{-1}$代替 $\Phi(\beta_0+\beta_1X_{1i}+\cdots+\beta_kX_{ki})$]。像 probit 模型一样，logit 系数的 MLE 也没有简单的计算公式，因此其对数似然函数也必须用数值代数方法进行最大化。



伪 R^2

伪 R^2 把估计模型的似然函数的值和没有变量 X 作为回归因子时的似然函数的值做比较。具体来说，伪 R^2 的值是：

的是 $Y_t=0$，而不是 Y_t^*。

当公式(9.21)使用观测到的支出 Y_t 代替 Y_t^* 进行估计时，所得出的OLS估计量是不一致的。Tobin通过使用 u_t 服从正态分布的这一额外假设来推导相应的似然函数，从而解决了这个问题，所得出的MLE被应用经济计量学家用于分析经济中的许多问题。为了纪念Tobin，公式(9.21)与正态误差假设一起被称为tobit回归模型。tobit模型是**删剪的回归模型**(censored regression model)的一个例子，之所以称其是删剪的，是因为因变量在某个切割点之上或之下的值已被删剪掉。

样本选择模型

在删剪的回归模型中，有关于购买者和非购买者的数据，从成人总体的简单随机抽样中获得的数据就是这样。然而，如果从销售税记录方面搜集数据，那么数据将只包含购买者，根本不会有非购买者的数据。在某一个门限值之上或者之下不可获得的观测值数据(只限于购买者的数据)被称为截尾数据。**截尾回归模型**(truncated regression model)是应用于因变量在某个切割点之上或者之下时数据的观测值得不到的数据中的一种回归模型。

截尾回归模型是样本选择模型的一个例子，在样本选择模型中，选择机制(个人由于买车而包含在样本中)与因变量的值(汽车的价格)有关系。如9.5节的一般兴趣框中所讨论的那样，估计样本选择模型的一种方法就是设计两个方程：一个是 Y_t^* 的方程，一个是不论 Y_t^* 是否被观测到的方程。然后模型的参数可以用极大似然法进行估计，或者在一个逐步法的程序中，首先估计选择方程，然后估计 Y_t^* 方程。进一步的讨论，请见Ruud(2000，第28章)或Greene(2000，第20.4节)。

计数数据

当因变量是一个计数的数时，例如一个消费者在一周内到饭店进餐的次数，**计数数据**(count data)就产生了。当这些数字很大时，该变量可被看做是近似连续的，但是当这些数字很小时，连续近似效果很差。用OLS估计的线性回归模型可被用于计数数据，即使计数的数很小。来自该回归的预测值，可被解释为以回归因子为条件的因变量的期望值。因此，当因变量是到饭店进餐的次数时，预测值1.7的意思是平均每周到饭店进餐1.7次。不过，就像在二元回归模型中一样，OLS没有利用计数数据的特殊结构，因此得到的是无意义的预测值，例如每周到饭店进餐次数为 -0.2。正如当因变量是二元变量时，probit模型和logit模型可以消除无意义的预测一样，对于计数数据，特殊模型也能做到，其中两个使用最广泛的模型就是泊松和负二项回归模型(Poisson and negative binomial regression model)。

有序的响应数据

当相互排斥的定性分类有一个正常的顺序时，例如获得一个高中学历、某种程度的大学教育(但没有毕业)，或者大学毕业，就出现了**有序的响应数据**(ordered response data)的现象。像计数数据一样，有序响应数据有一个自然的顺序，但是与计数数据又不同，有序的响应数据没有自然的数值。



由于有序的响应数据没有自然的数值，所以OLS是不适用的，而是经常使用一个被称为**有序probit模型**(ordered probit model)的probit模型的推广来分析有序数据，在这个模型中，以自变量(如父母的收入)为条件的每个结果(如大学教育)的概率使用累积正态分布来建立模型。

离散选择数据

离散选择(discrete choice)变量或多个选择(multiple choice)变量可以取多个无序的定性值。经济学中一个例子是,一位通勤者所选择的交通方式可能是乘地铁、坐公交车、自己开车,或者自己支配的其他方式(如步行、骑车)。如果我们要分析这些选择,那么因变量会有四种可能的结果(地铁、公交车、汽车、人力)。无论以何种自然的方式排列,这些结果都不是有序的。相反,这些结果是不同的可供选择的定性方案中的一种选择。

经济计量学的任务就是在给定不同回归因子的条件下,例如个体特征(通勤者住处离地铁站有多远)和每个选项的特征(地铁票价),建立选择不同选项的概率模型。如9.5节的一般兴趣框中所讨论的,离散选择数据的分析模型可以用效用最大化原理来推导。个体选择概率可以用probit模型或者logit模型形式来表达,而这些模型被称为多项probit(multinomial probit)和多项logit(multinomial logit)回归模型。

相关性可能有不同的根源，包括省略变量、变量误差（回归因子中的测度误差），或联立因果关系（这时的因果关系既从 Y 到 X 即"向后"传导，也从 X 到 Y 即"向前"传导）。不论 X 和 u 之间的相关性的来源如何，如果存在一个有效的工具变量 Z，那么就可以使用该工具变量的估计量来估计 X 的单位变化对 Y 的影响。

10.1.1 IV 模型和假设

联系因变量 Y_i 和回归因子 X_i 的总体回归模型是：

$$Y_i = \beta_0 + \beta_1 X_i + u_i,\ i = 1, \cdots, n \tag{10.1}$$

和通常一样，u_i 是代表决定 Y_i 的遗漏因素的误差项。如果 X_i 与 u_i 相关，那么 OLS 估计量就是不一致的。工具变量估计使用一个额外的"工具"变量 Z 来分离出与 u_i 无关的那部分 X 的变化。

内生性和外生性。工具变量回归有一些专门的术语，这些术语将与总体误差项 u 相关的变量和不相关的变量区分开来。与误差项相关的变量被称为**内生**（endogenous）变量，而与误差项无关的变量被称为**外生**（exogenous）变量。这些术语的历史渊源可追溯到含有多个方程的模型，在这些模型中，"内生"变量是在模型内部决定的，而"外生"变量是在模型外部决定的。例如，7.2 节考虑了这种可能性，如果由于政治干预和增加资助导致低的考试成绩会引起学生—教师比的下降，那么学生—教师比和考试成绩之间就互为因果关系。在数学上，这可被表示为一个有两个联立方程公式（7.3）和公式（7.4）的系统，每个方程对应一个因果关系。如 7.2 节中所讨论的，因为考试成绩和学生—教师比都是在模型内决定的，所以它们都与总体误差项 u 相关，也就是说，在这个例子中，两个变量都是内生的。相反，在模型外部决定的外生变量与 u 无关。

有效工具变量的两个条件。一个有效的工具变量（"工具"）必须满足两个条件，即所谓的**工具变量相关性**（instrument relevance）和**工具变量外生性**（instrument exogeneity）：

1. 工具变量相关性：$\text{corr}(Z_i, X_i) \neq 0$。
2. 工具变量外生性：$\text{corr}(Z_i, u_i) = 0$。

如果一个工具变量是相关的，那么该工具变量的变化与 X_i 的变化相关。此外，如果该工具变量是外生的，那么由该工具变量所捕捉到的 X_i 的那部分变化就是外生的。因此，一个相关的而且外生的工具变量能够捕捉 X_i 的外生变化，这个外生变化反过来能被用来估计总体系数 β_1。

对于工具变量回归而言，一个有效工具变量的这两个条件是至关重要的，在整章中我们都会重复这两个条件（它们对多个回归因子和多个工具变量的情形仍然是很重要的）。

10.1.2 两阶段最小二乘估计量

如果工具变量 Z 满足工具变量相关性和外生性的条件，那么可以使用被称为**两阶段最小二乘**（two stage least squares，简写为 TSLS）的 IV 估计量来估计系数 β_1。顾名思义，两阶段最小二乘估计量是按两个阶段计算的。第一阶段将 X 分解成两部分：一个是可能与回归误差项相关的有问题的部分；另一个是与回归误差项无关的没有问题的部分。第二阶段使用这个没有问题的部分来估计 β_1。

第一阶段从联系 X 和 Z 的总体回归开始：

$$X_i = \pi_0 + \pi_1 Z_i + v_i \tag{10.2}$$

其中，π_0 是截距，π_1 是斜率，v_i 是误差项。这个回归提供了我们对 X_i 所需要的分解。一部



分是 $\pi_0+\pi_1 Z_i$，它是可以被 Z_i 预测的 X_i 中的部分。由于 Z_i 是外生的，因此，X_i 中的这个部分与公式(10.1)中的误差项 u_i 无关。X_i 中的另一部分就是 v_i，它是与 u_i 相关的 X_i 中有问题的部分。

TSLS 背后的思想就是，使用无问题的 X_i 的部分，即 $\pi_0+\pi_1 Z_i$，忽略 v_i。惟一的复杂之处就是 π_0 和 π_1 的值都是未知的，因此不能计算 $\pi_0+\pi_1 Z_i$。所以，TSLS 的第一阶段将 OLS 应用于公式(10.2)，并且使用 OLS 回归的预测值 $\hat{X}_i=\hat{\pi}_0+\hat{\pi}_1 Z_i$，其中，$\hat{\pi}_0$ 和 $\hat{\pi}_1$ 是 OLS 估计值。

TSLS 的第二阶段很容易，用 Y_i 对 $\hat{X}_i$ 进行 OLS 回归。从第二阶段回归中得出的相应的估计量就是 TSLS 估计量，即 $\hat{\beta}_0^{TSLS}$ 和 $\hat{\beta}_1^{TSLS}$。

10.1.3 为什么 IV 回归会起作用

对于为什么 IV 回归解决了 X_i 和 u_i 之间的相关性问题，下面两个例子提供了一些直观解释。

例 1：Philip Wright 的问题。工具变量估计方法最早发表在 1928 年由 Philip G. Wright (Wright，1928)所编写的一本书的附录中，但是人们认为这个附录是他的儿子 Sewall Wright——一个重要的统计学家所写的，或者是他俩合写的。Philip Wright 关心他那个时代的一个重要的经济问题：如何制定动植物油和油脂的进口关税(对进口商品征收的一种税)，例如黄油和豆油。在 20 世纪 20 年代，进口关税是美国税收的主要来源。理解关税的经济效应的关键就是要有对该商品供求曲线的定量估计值。回想一下，供给弹性就是价格增加 1% 所引起的供给量的百分比变化，而需求弹性就是价格增加 1% 所引起的需求量的百分比变化。Philip Wright 需要的就是这些供给和需求弹性的估计值。

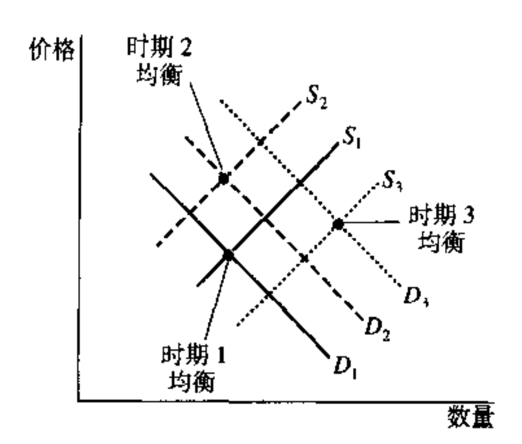
为了使问题具体化，考虑估计黄油需求弹性的问题。回想一下重要概念 6.2，联系 $\ln(Y_i)$ 和 $\ln(X_i)$ 的线性方程的系数，其含义就是 Y 关于 X 的弹性。在 Wright 父子的问题中，意味着需求方程可以这样写：

$$\ln(Q_i^{butter})=\beta_0+\beta_1\ln(P_i^{butter})+u_i \qquad (10.3)$$

其中，Q_i^{butter} 是黄油消费量的第 i 个观测值，P_i^{butter} 是黄油的价格，u_i 代表影响需求的其他因素，例如收入和消费者嗜好。在公式(10.3)中，黄油价格每增加 1% 就会引起需求量变化百分之 β_1，因此，β_1 就是需求弹性。

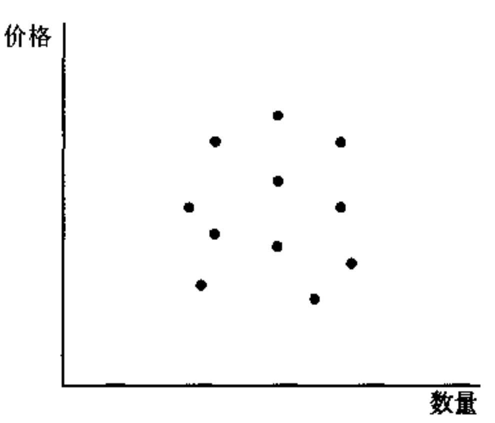
Wright 父子拥有 1912 年到 1922 年美国黄油年度总消费量及其年度平均价格的数据。通过将 OLS 应用于公式(10.3)，将会很容易使用这些数据来估计需求弹性，但是他们有个重要的看法，即由于供给和需求之间的相互作用，回归因子 $\ln(P_i^{butter})$ 可能与误差项相关。

为了弄清这一点，请看图 10—1(a)。该图显示了三个不同年份里黄油的市场需求和供给曲线。第一个时期的需求和供给曲线分别用 D_1 和 S_1 表示，而第一个时期的均衡价格和均衡数量是由它们的交点决定的。在第二年，需求从 D_1 上升到 D_2(比如，由于收入的增加)，而供给从 S_1 下降到 S_2(比如，由于黄油生产成本的上升)，均衡价格和均衡数量是由新的供给和需求曲线的交点决定的。在第三年，影响需求和供给的因素又发生了变化，需求增加到 D_3，供给增加到 S_3，并决定了新的均衡价格和均衡数量。图 10—1(b)显示了这三年以及接下来八年的均衡数量和均衡价格对，其中在每一年里，供给和需求曲线都受到了除价格之外的影响市场供给和需求变化的因素的影响。这个散点图就像 Wright 父子在描绘他们的数据时所看到的散点图一样。就如他们所推理的一样，用 OLS 拟合这些点的直线既不能估计需求曲线，也不能估计供给曲线，因为这些点是由需求和供给的变化共同决定的。



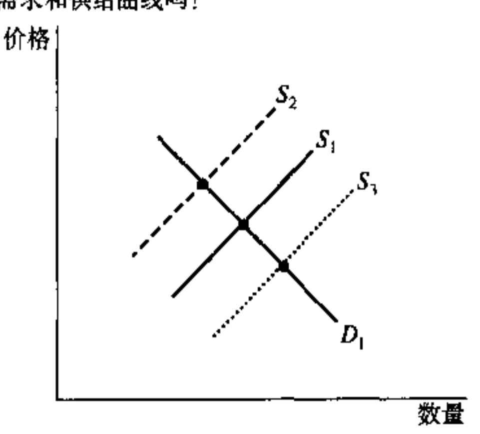
(a) 3 个时期的需求和供给

注:价格和数量是由供给曲线和需求曲线的交点决定的。第一个时期的均衡是由需求曲线 D_1 和供给曲线 S_1 的交点决定的,第二个时期的均衡是由 D_2 和 S_2 的交点决定的,而第三个时期的均衡是由 D_3 和 S_3 的交点决定的。



(b) 11 个时期的均衡价格和均衡数量

注:这个散点图显示了 11 个不同时期里的均衡价格和均衡数量,隐藏了需求和供给曲线。你能够根据这个散点图中的点确定需求和供给曲线吗?



(c) 在只有供给曲线移动时的均衡价格和均衡数量

注:当供给曲线从 S_1 移动到 S_2 再到 S_3 而需求曲线保留在 D_1 时,均衡价格和均衡数量沿着需求曲线移动。

图 10—1　黄油的供给、需求与价格



Wright 父子意识到，处理这个问题的方法就是要找到使供给曲线移动而又不影响需求曲线变化的某一第三个变量。图 10—1(c)显示了当这样一个变量使供给曲线发生移动，但需求曲线保持不变时的情景。现在所有的均衡价格和数量对都在一条稳定的需求曲线上，并且这条需求曲线的斜率很容易被估计。在 Wright 父子问题的工具变量表达式中，第三个变量——工具变量——与价格相关(它使供给曲线移动，这导致了价格的变化)，但与 u 无关(需求曲线保持稳定不变)。Wright 父子研究了几个可能的工具变量，其中一个是天气。例如，奶牛场地区低于平均水平的降雨量可能会影响牧草生长，这样会减少给定价格下的黄油产量(它将会使供给曲线向左移动，提高均衡价格)，所以奶牛场地区的降雨量满足工具变量相关性这一条件，但是奶牛场地区的降雨量不应该对黄油的需求有直接影响，因此，奶牛场地区降雨量与 u_i 之间的相关系数将会是 0。也就是说，奶牛场地区降雨量还满足工具变量外生性这一条件。

例 2：估计班级规模对考试成绩的影响。尽管控制了学生和地区特征，但第 2 部分所给出的班级规模对考试成绩影响的估计值，仍可能含有由不可测度的变量，例如校外学习机会或教师质量等，引起遗漏变量偏差。如果得不到这些变量的数据，那么这个遗漏变量偏差就不能通过在多元回归中引入这些变量来解决。

工具变量回归就为这个问题提供了一个可供选择的解决方法。考虑下面的假设例子：夏天由于地震，加利福尼亚州的一些学校被迫停课维修，接近震中的地区受到最严重的影响。一些停课的地区必须把学生挤在一起，临时扩大班级规模。这意味着与震中的距离满足工具变量相关性的条件，因为它与班级规模相关。但是，如果与震中距离不与任何影响学生成绩的其他因素相关(例如，学生是否仍然学英语)，那么它将是外生的，因为它与误差项不相关。因而，到震中的距离这个工具变量，就可以被用来防止遗漏变量偏差的发生和估计班级规模对考试成绩的影响。

10.1.4 TSLS 估计量的抽样分布

在小样本中，TSLS 估计量的精确分布很复杂。不过，像 OLS 估计量一样，在大样本条件下它的分布是简单的：TSLS 估计量是一致的且服从正态分布。

TSLS 估计量的公式。虽然 TSLS 的两个阶段使估计量看上去显得复杂，但是，如果只有一个 X 和一个工具变量 Z 时(如本节我们所假设的)，TSLS 估计量的公式就很简单。假设 s_{ZY}是 Z 和 Y 之间的样本协方差，假设 s_{ZX}是 Z 和 X 之间的样本协方差，如附录 10.2 所示，含单个工具变量的 TSLS 估计量是：

$$\hat{\beta}_1^{TSLS}=\frac{s_{ZY}}{s_{ZX}} \tag{10.4}$$

也就是说，β_1 的 TSLS 估计量就是 Z 和 Y 之间的样本协方差与 Z 和 Y 之间的样本协方差之比。

当样本容量很大时，$\hat{\beta}_1^{TSLS}$ 的抽样分布。公式(10.4)可被用来证明 $\hat{\beta}_1^{TSLS}$ 是一致的，而且在大样本条件下服从正态分布。这里只是把争论要点总结一下，而数学细节则在附录 10.3 中给出。



$\hat{\beta}_1^{TSLS}$ 是一致的这一论点，把 Z_i 是相关的和外生的假设同样本协方差和总体协方差相一致这个结论结合起来了。因为在公式(10.1)中 $Y_i=\beta_0+\beta_1X_i+u_i$，所以：

$$\mathrm{cov}(Z_i,Y_i)=\mathrm{cov}(Z_i,\beta_0+\beta_1X_i+u_i)=\beta_1\mathrm{cov}(Z_i,X_i)+\mathrm{cov}(Z_i,u_i) \tag{10.5}$$

其中，第二个等式是根据协方差的性质(公式(2.33))得到的。根据工具变量外生性假设，

$\text{cov}(Z_i,u_i)=0$，而根据工具变量相关性假设，$\text{cov}(Z_i,X_i)\neq0$。因而，如果工具变量是有效的，那么：

$$\beta_1=\frac{\text{cov}(Z_i,Y_i)}{\text{cov}(Z_i,X_i)} \tag{10.6}$$

这就是说，总体系数 β_1 就是 Z 和 Y 之间的总体协方差与 Z 和 X 之间的总体协方差之比。

如 3.6 节中所讨论的，样本协方差是总体协方差的一致估计量，也就是说，$s_{ZY}\xrightarrow{P}\text{cov}(Z_i,Y_i)$，$s_{ZX}\xrightarrow{P}\text{cov}(Z_i,X_i)$。由公式(10.4)和公式(10.6)得出的 TSLS 估计量是一致的：

$$\hat{\beta}_1^{TSLS}=\frac{s_{ZY}}{s_{ZX}}\xrightarrow{P}\frac{\text{cov}(Z_i,Y_i)}{\text{cov}(Z_i,X_i)}=\beta_1 \tag{10.7}$$

公式(10.4)也能被用来证明在大样本条件下，$\hat{\beta}_1^{TSLS}$ 的抽样分布是正态分布，理由与我们已考虑过的其他每一个最小二乘估计量相同：TSLS 估计量是随机变量的一个平均值，而在样本规模很大时，中心极限定理告诉我们，随机变量的平均值服从正态分布。具体来说，公式(10.4)中的 $\hat{\beta}_1^{TSLS}$ 表达式的分子 $s_{ZY}=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})(Y_i-\overline{Y})$ 是 $(Z_i-\overline{Z})(Y_i-\overline{Y})$ 的一个平均值。在附录 10.3 中所介绍的有关代数知识表明，由于求这个平均值，中心极限定理隐含在大样本条件下，$\hat{\beta}_1^{TSLS}$ 有个逼近于 $N(\beta_1,\sigma^2_{\hat{\beta}_1^{TSLS}})$ 的抽样分布，其中：

$$\sigma^2_{\hat{\beta}_1^{TSLS}}=\frac{1}{n}\frac{\text{var}[(Z_i-\mu_Z)u_i]}{[\text{cov}(Z_i-X_i)]^2} \tag{10.8}$$

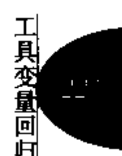
使用大样本分布的统计推断。方差 $\sigma^2_{\hat{\beta}_1^{TSLS}}$ 可通过估计公式(10.8)中出现的方差和协方差来估计，而 $\sigma^2_{\hat{\beta}_1^{TSLS}}$ 估计值的平方根就是 IV 估计量的标准误。在经济计量学软件包里，TSLS 回归命令会自动计算这个标准误。因为在大样本条件下，$\hat{\beta}_1^{TSLS}$ 服从正态分布，所以关于 β_1 的假设检验可通过计算 t 统计量来进行，而 95% 的大样本置信区间由 $\hat{\beta}_1^{TSLS}\pm1.96SE(\hat{\beta}_1^{TSLS})$ 给出。

10.1.5 在香烟需求案例中的应用

Wright 父子对黄油的需求弹性感兴趣，但是在今天，其他商品，例如香烟，在公共政策争论方面显得更重要了。为了减少由吸烟引起的疾病和死亡以及那些疾病强加在其他社会成员身上的成本或外部性，所寻求的一个工具就是征收很重的香烟税，以使现有的吸烟者戒烟并劝阻潜在的新吸烟者不去吸烟。但是确切地说，需要增加多大的税收才会削弱香烟的消费？例如，要使香烟消费减少 20%，税后的香烟价格应该是多少？

对这个问题的回答依赖于香烟的需求弹性。如果弹性是 -1，那么价格增加 20% 就能达到使消费减少 20% 的目标；如果弹性是 -0.5，那么价格必须增加 40%，才能使消费减少 20%。当然理论上我们并不知道香烟的需求弹性是多少，我们必须从价格和消费的数据中估计它。但是，与黄油一样，由于供给和需求之间的相互作用，香烟的需求弹性不能被对数消费量对对数价格的 OLS 回归函数一致地估计。因此我们使用 TSLS 来估计香烟的需求弹性，数据采用 1985 年到 1995 年美国 48 个大陆州的年度数据（附录 10.1 中描述了该数据）。目前，所有结论都是 1995 年各州截面数据的结论，更早年份的面板数据的结论在 10.4 节中给出。

工具变量 $SaleTax_i$，即香烟税，是出现在总销售税中以每包的美元数来测度（以实际美元测度的，用消费者价格指数进行缩减）的部分。香烟消费量（$Q_i^{cigarettes}$）是该州内人均销售



内含外生变量(included exogenous variables),我们将其标记为 W;以及工具变量 Z。一般地说,可能有多个内生回归因子(X)、多个外生回归因子(W)和多个工具变量(Z)。

为了使 IV 回归成为可能,至少要有与内生回归因子(X)一样多的工具变量(Z)。在 10.1 节,只有一个内生变量和一个工具变量。对这个单一的内生回归因子而言,必须至少有一个工具变量。如果没有工具变量,我们就不能计算工具变量的估计量,不会有 TSLS 的第一阶段回归。

重要概念 10.1

一般的工具变量回归模型和术语

一般的 IV 回归模型是:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \cdots + \beta_k X_{ki} + \beta_{k+1} W_{1i} + \cdots + \beta_{k+r} W_{ri} + u_i, i = 1, 2, \cdots, n \tag{10.12}$$

其中:

- Y_i 是因变量;
- u_i 是误差项,它代表测度误差和/或遗漏因素;
- $X_{1i}, \cdots, X_{ki}$ 是 k 个内生回归因子,它们可能与 u_i 相关;
- $W_{1i}, \cdots, W_{ri}$ 是 r 个内含的外生回归因子,它们与 u_i 不相关;
- $\beta_0, \beta_1, \cdots, \beta_{k+r}$ 是未知的回归系数;
- $Z_{1i}, \cdots, Z_{mi}$ 是 m 个工具变量。

如果工具变量比内生回归因子多($m > k$),那么系数被过度识别;如果 $m < k$,那么系数被识别得不足;如果 $m = k$,那么系数被恰好识别。IV 回归模型的估计要求恰好识别或过度识别。

工具变量的个数和内生回归因子的个数之间的关系足以重要到需要有其自己的术语。如果工具变量的个数(m)等于内生回归因子的个数(k),即 $m = k$,那么回归系数被称为是恰好识别(exactly identified)的;如果工具变量的个数(m)超过内生回归因子的个数(k),即 $m > k$,那么回归系数被称为是过度识别(overidentified)的;如果工具变量的个数(m)少于内生回归因子的个数(k),即 $m < k$,那么回归系数被称为是不足识别(underidentified)的。如果要用 IV 回归估计系数,那么系数一定是恰好识别的,或者是过度识别的。

一般的 IV 回归模型及其术语在重要概念 10.1 中做了总结。

10.2.1 一般 IV 模型中的 TSLS

含有单个内生回归因子的 TSLS。当含有单个内生回归因子 X 和一些额外的内含外生变量时,相应的方程是:

$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i \tag{10.13}$$

其中,和以前一样,X_i 可能与误差项相关,但 $W_{1i}, \cdots, W_{ri}$ 与误差项不相关。

TSLS 的总体第一阶段回归将 X 与外生变量,即 W 和工具变量 Z 联系在一起:

$$X_i = \pi_0 + \pi_1 Z_{1i} + \cdots + \pi_m Z_{mi} + \pi_{m+1} W_{1i} + \cdots + \pi_{m+r} W_{ri} + v_i \tag{10.14}$$

其中,$\pi_0, \pi_1, \cdots, \pi_{m+r}$ 是未知的回归系数,v_i 是误差项。

公式(10.14)有时被称为 X 的简约式(reduced form)方程,它将内生变量 X 与所有可获得的外生变量联系在一起,外生变量既包括回归方程中的变量(W),也包括工具变量(Z)。

在 TSLS 的第一阶段,公式(10.14)中的未知系数由 OLS 来估计,而且这个回归的预测



值是 $\hat{X}_1,\cdots,\hat{X}_n$。

在 TSLS 的第二阶段，公式(10.13)使用 OLS 进行估计，只是 X_i 用第一阶段它的预测值做了替代，也就是说，用 Y_i 对 $\hat{X}_i, W_{1i},\cdots,W_{ri}$ 进行 OLS 回归。相应所得的 $\beta_0,\beta_1,\cdots,\beta_{1+r}$ 的估计量就是 TSLS 估计量。

推广到多个内生回归因子。当有多个内生回归因子 $X_{1i},\cdots,X_{ki}$ 时，除了每个内生回归因子要求有它自己的第一阶段回归之外，TSLS 的算法是相似的。这些第一阶段回归中的每一个都有与公式(10.14)相同的形式，也就是说，因变量是 X 变量中的一个，而回归因子是所有的工具变量(Z)和所有的内含外生变量(W)。这些第一阶段回归共同生成了每个内生回归因子的预测值。

在 TSLS 的第二阶段，公式(10.12)用 OLS 来估计，只是内生回归因子(X)被它们各自的预测值($\hat{X}$)所代替，所得出的相应的 $\beta_0,\beta_1,\cdots,\beta_{1+r}$ 的估计量就是 TSLS 估计量。

实践中，现代经济计量软件的 TSLS 估计命令会自动运行 TSLS 的两个阶段回归。一般的 TSLS 估计量在重要概念 10.2 中做了总结。

重要概念 10.2

两阶段最小二乘

在公式(10.12)一般的 IV 回归模型中，含有多个工具变量的 TSLS 估计量是分两个阶段来计算的。

1. 第一阶段回归(first-stage regression(s))：用 X_{1i} 对工具变量($Z_{1i},\cdots,Z_{mi}$)和内含的外生变量($W_{1i},\cdots,W_{ri}$)进行 OLS 回归。计算这个回归的预测值，被称为 $\hat{X}_{1i}$。对所有内生回归因子 $X_{2i},\cdots,X_{ki}$ 重复以上步骤，这样就计算出了预测值 $\hat{X}_{1i},\cdots,\hat{X}_{ki}$。

2. 第二阶段回归(second-stage regression)：用 Y_i 对内生变量预测值($\hat{X}_{1i},\cdots,\hat{X}_{ki}$)和内含的外生变量($W_{1i},\cdots,W_{ri}$)进行 OLS 回归。TSLS 的估计量 $\hat{\beta}_0^{TSLS},\cdots,\hat{\beta}_{k+r}^{TSLS}$ 就是第二阶段回归的估计量。

在实际中，这两个阶段的回归可由现代经济计量软件中的 TSLS 估计命令自动完成。

10.2.2 一般 IV 模型中的工具变量相关性和外生性

对于一般的 IV 回归模型，工具变量相关性和外生性的条件需要进行修改。

当有一个内含外生变量但有多个工具变量时，工具变量相关性的条件是，对给定的 W，至少有一个 Z 对预测 X 是有用的。当有多个内含的外生变量时，这个条件更为复杂，因为我们必须排除总体回归中的完全多重共线性。直观上，当有多个内含的外生变量时，工具变量必须提供足够多的关于这些变量外生变化的信息以区分它们各自对 Y 的影响。

工具变量外生性条件的一般陈述是，每个工具变量必须与误差项 u_i 不相关。关于有效工具变量的一般条件在重要概念 10.3 中给出。



重要概念 10.3

有效工具变量的两个条件

一个含 m 个工具变量 $Z_{1i},\cdots,Z_{mi}$ 的组合必须满足下列两个条件才是有效的：

估计所给出的标准误并不正确，因为它们没有认识到它是两阶段过程的第二个阶段。具体来说，第二阶段 OLS 标准误没能对第二阶段的回归使用了内含内生变量的预测值这一事实进行相应的调整。进行必要调整的标准误公式被嵌入到(且被自动使用)经济计量软件的 TSLS 回归命令中。因此，如果你使用专门的 TSLS 回归软件，那么在实际中这个问题就不成为问题了。第二，像通常一样，误差项 u 可能是异方差的，因此使用标准误的异方差稳健形式是非常重要的，它与多元回归模型的 OLS 估计量使用异方差稳健的标准误是重要的理由完全相同。

10.2.5 在香烟需求案例中的应用

在 10.1 节中，我们利用 1995 年美国 48 个州的年度消费数据，使用含有单个回归因子(每包香烟实际价格的对数)和单个工具变量(每包香烟的实际销售税)的 TSLS 估计了香烟的需求弹性。然而，收入也影响需求，因此它也是总体回归误差项的一部分。就如 10.1 节中所讨论的，如果州的销售税与该州的收入有关，那么它与香烟需求方程误差项中的某个变量相关，这就违背了工具变量外生性条件。如果是这样的话，10.1 节中的 IV 估计量就是不一致的，也就是说，IV 回归受省略变量偏差的影响。为了解决这个问题，我们需要在回归中引入收入因子。

因此，我们考虑一个可供选择的设定，在这个设定中，收入的对数被包含在需求方程中。根据重要概念 10.1 的术语说明，因变量 Y 是消费量的对数，即 $\ln(Q_i^{cigarettes})$；内生回归因子 X 是实际价格的对数，即 $\ln(P_i^{cigarettes})$；内含外生变量 W 是实际人均州收入的对数 $\ln(Inc_i)$；工具变量 Z 是每包香烟的实际销售税 $SalesTax_i$。TSLS 估计值和(异方差稳健的)标准误是：

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(1.26)}{9.43} - \underset{(0.37)}{1.14}\ln(P_i^{cigarettes}) + \underset{(0.31)}{0.21}\ln(Inc_i) \tag{10.15}$$

这个回归使用单个工具变量 $SalesTax_i$，但是实际上可获得另一个候选的工具变量。除了总销售税以外，州政府征收只适用于香烟和烟草产品的特产税。这些香烟特产税($CigTax_i$)构成了另一个可能的工具变量。香烟特产税会使消费者支付的香烟价格增加，因此，可证明它满足工具变量相关性的条件。如果它与州香烟需求方程中的误差项不相关，那么它就是一个外生的工具变量。

手中有这个额外的工具变量，我们现在就有了两个工具变量：每包香烟的实际销售税和每包香烟的实际州特产税。由于有两个工具变量和一个内生回归因子，所以需求弹性被过度识别，也就是说，工具变量的个数($SalesTax_i$ 和 $CigTax_i$，因此 $m=2$)超过了内含内生变量的个数($P_i^{cigarettes}$，因此 $k=1$)。我们可以使用 TSLS 估计需求弹性，其中第一阶段中的回归因子是内含外生变量 $\ln(Inc_i)$ 和两个工具变量。

使用两个工具变量 $SalesTax_i$ 和 $CigTax_i$ 所得出的回归函数的 TSLS 估计值是：

$$\widehat{\ln(Q_i^{cigarettes})} = \underset{(0.96)}{9.89} - \underset{(0.25)}{1.28}\ln(P_i^{cigarettes}) + \underset{(0.25)}{0.28}\ln(Inc_i) \tag{10.16}$$

比较公式(10.15)和公式(10.16)：在公式(10.16)中，所估计的价格弹性的标准误要小 1/3(公式(10.16)中的 0.25 比公式(10.15)中的 0.37)。公式(10.16)中标准误比较小的原因是这个估计值使用了比公式(10.15)更多的信息：在公式(10.15)中，只使用了一个工具变量(销售税)，而在公式(10.16)中，使用了两个工具变量(销售税和香烟特产税)。使用两个工具变量比只使用一个工具变量解释了更多的香烟价格变化，而这又被反映在所估计的需求弹性有较小的标准误差上。



这些使用两个工具变量的估计值更可信吗？最终，可信性依赖于这个工具变量组——这里是两个税收变量——是否能够合理地满足有效工具变量的两个条件。因此，我们评估这些工具变量是否有效是至关重要的，现在我们就来讨论这个主题。

10.3 检查工具变量的有效性

在一个特定的应用中，工具变量回归是否有用取决于工具变量是否有效，无效的工具变量生成无意义的结论。因此，在一个特定的应用中，评价一组给定的工具变量是否有效是非常必要的。

10.3.1 假设1：工具变量相关性

在IV回归中，工具变量相关性条件的作用是微妙的。理解工具变量相关性的一种方法是，它起着类似样本容量的作用：工具变量的相关性越强——即 X 的变差被工具变量所解释的越多——在IV回归中可获得的信息越多。一个工具变量的相关程度越高，生成的估计量越准确，正如越大的样本会生成越准确的估计量一样。此外，使用TSLS的统计推断是根据TSLS估计量服从正态抽样分布的假设来预测的，但是根据中心极限定理，在大样本条件下（小样本条件下不一定成立），正态分布是个很好的逼近。如果一个具有较高相关性的工具变量就像是拥有较大的样本容量的特征一样，那么这意味着，由于正态分布为TSLS估计量的抽样分布提供了一个很好的逼近，因此这些工具变量不应该只是相关的，而应该是高度相关的。

几乎没有解释 X 变化的工具变量被称为**弱工具变量**（weak instruments）。在香烟的那个例子中，该州到香烟制造厂之间的距离可以认为是个弱工具变量：虽然较长距离会增加运输成本（这样供给曲线向里移动，使均衡价格上升），但是香烟的重量很轻，因此运输成本只是香烟价格中很小的一部分。这样，价格变差中由运输成本所解释的量，也即由该州到香烟制造厂的距离所解释的量，可能会非常小。

本节讨论为什么弱工具变量是个问题，如何识别弱工具变量，以及如果存在弱工具变量，你应该做些什么。在此，始终假设工具变量是外生的。

为什么弱工具变量是个问题？如果工具变量是弱的，即使样本容量很大，正态分布为TSLS估计量的抽样分布也只能提供一个比较差的逼近。这样，进行统计推断的通常方法就没有了理论上的依据，即使在大样本条件下。事实上，如果工具变量是弱的，那么TSLS估计量可能会产生严重的偏差，被构造为TSLS估计量±1.96倍标准误的95%的置信区间可能包含了远小于95%的机会的系数的真实值。简而言之，如果工具变量是弱的，那么TSLS就不再可靠了。

为了弄明白TSLS估计量抽样分布在大样本条件下的正态逼近也有问题，考虑一个特殊的例子，即10.1节中所考虑的含有单个内含内生变量、单个工具变量以及无内含外生回归因子的例子。如果该工具变量是有效的，那么 $\hat{\beta}_1^{TSLS}$ 就是一致的，因为样本协方差 s_{ZY} 和 s_{ZX} 是一致的，也就是说，$\hat{\beta}_1^{TSLS}=s_{ZY}/s_{ZX}\xrightarrow{P}\mathrm{cov}(Z_i,Y_i)/\mathrm{cov}(Z_i,X_i)=\beta_1$（表达式(10.7)）。但是，现在假设工具变量不仅是弱的，而且还是不相关的，那么有 $\mathrm{cov}(Z_i,X_i)=0$。这样，$s_{ZX}\xrightarrow{P}\mathrm{cov}(Z_i,X_i)=0$。因而，按照字面理解，右边极限 $\mathrm{cov}(Z_i,Y_i)/\mathrm{cov}(Z_i,X_i)$ 的分母是0！显然，当工具变量相关性不成立时，$\hat{\beta}_1^{TSLS}$ 是一致的论点就崩溃了。如附录10.4所证明的，这



个论点的崩溃导致了TSLS估计量有个非正态的抽样分布，即使样本容量非常大。事实上，当工具变量不相关时，$\hat{\beta}_1^{TSLS}$ 的大样本分布并不是正态随机变量的大样本分布，换句话说，它是两个正态随机变量之比率的分布！

然而在实际中不太可能会遇到这种完全不相关的工具变量的情形，这样就提出了一个问题：在实际中，工具变量必须达到多大程度的相关才能为正态分布提供一个很好的逼近呢？在一般的IV回归模型中，这个问题的答案是复杂的。然而，幸运的是，对于实际中最一般的情形——单个内生回归因子的情形，存在一个可获得的简单的经验规则。

重要概念10.5

识别弱工具变量的一个经验规则

第一阶段的 F 统计量是检验两阶段最小二乘法第一阶段中工具变量 $Z_{1i},\cdots,Z_{mi}$ 的系数都等于0这一假设的 F 统计量。当有一个内生回归因子时，第一阶段 F 统计量小于10则表明，该工具变量是弱工具变量，在此情形下，TSLS估计量是有偏的（即使在大样本条件下），而且TSLS的 t 统计量和置信区间都是不可靠的。

当有一个内生回归因子时，弱工具变量的检验。当只有一个内生回归因子时，检验弱工具变量的一种方法是，计算检验TSLS第一阶段回归中工具变量系数都等于0这一假设的 F 统计量。这个第一阶段 F 统计量（first stage F-statistic）提供了工具变量中所包含的信息内容的测度：信息内容越多，F 统计量的期望值就越大。一个简单的经验规则是，如果第一阶段的 F 统计量超过10，那么你就不必担心弱工具变量了（为什么是10呢？见附录10.4。）。这个内容在重要概念10.5中总结。

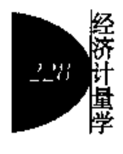
如果检验出弱工具变量，那么我该做些什么？依情况而定。如果你有许多工具变量，那么有些工具变量可能比其他变量弱。如果你有极少数的强工具变量但有较多的弱工具变量，那么你最好是舍弃最弱的工具变量，而将相关性最强的子集应用于你的TSLS分析中。当你舍弃弱工具变量时，你的TSLS标准误可能会增加，但是要记住你最初的标准误是毫无意义的。

然而，如果你只有几个工具变量或者如果系数被恰好识别，那么舍弃弱工具变量将不会有帮助。在此情形下，有两个办法：寻找另外的更强的工具变量，或者使用一些为使用弱工具变量而特别设计的高级分析工具。第一个办法要求熟悉手中的问题，而且可能需要重新设计数据集和该项实证研究的性质。第二个办法要求采用比TSLS对弱工具变量更不敏感的程序和方法。一个比TSLS对弱工具变量更不敏感的估计量是有限信息极大似然估计量（LIML），见Hayashi（2000，第8.6节）或Greene（2000，第16章）。当工具变量是弱工具变量时，发展更可靠的方法和程序是当前学术界一个活跃的研究领域。

一般兴趣框

一个引起惊慌的回归

一种估计多上一年学所引致收入百分比增加（“教育的回报”）的方法，是使用个人数据中收入的对数对受教育年数进行回归。但是如果有能力的个人既在劳动力市场中很成功，上学时间也很长（也许是因为这对他们更容易），那么受教育年数就会与“天赋”这一遗漏变量相关，进而教育回报的OLS估计量将是有偏的。由于先天的能力特别难测度，因此不能被用做回归因子，于是一些劳动经济学家求助于IV回归来估计教育的回报。但是，在收入



的回归中,什么变量与受教育年数相关而又与误差项不相关呢？也就是说,什么样的变量是有效的工具变量？

劳动经济学家 Joshua Angrist 和 Alan Krueger 建议是你的生日。由于义务教育法,他们推理认为你的生日与受教育年数相关:如果法律要求你上学上到 16 岁为止,并且你在 1 月份满 16 岁,这时你在十年级,那么你可能会退学。但是,如果你在 7 月份满 16 岁,那时你就已读完了十年级。如果是这样的话,你的生日满足工具变量相关性条件,但是在 1 月出生或在 7 月出生应该对你的收入没有直接影响(除了通过受教育年数),因此,你的生日又满足了工具变量的外生性条件。他们通过使用某个人出生所在的季度(三个月的时期)作为工具变量来实现上述的思路。他们使用了一个很大的来自于美国人口普查的数据样本(他们的回归至少有 329 000 个观测值!),而且他们控制了其他变量,例如工人年龄。

但是,另一位劳动经济学家 John Bound 对此表示怀疑。他知道弱工具变量会使 TSLS 不可靠,并且担心虽然样本容量特别大,但是在变量的一些设定中,出生的季度可能是个弱工具变量。因此,当 Bound 和 Krueger 下次进餐碰面时,争论一定转到 Angrist - Krueger 工具变量是否是弱工具变量的问题上。Krueger 认为不是,并且建议一个查明真相的创造性方法:为什么不使用一个确实不相关的工具变量——用计算机随机生成的假的出生季度代替每个人的实际出生季度——重新进行回归,并且把使用实际工具变量和使用假的工具变量的结果进行比较呢？他们发现的结果令人惊讶:不论你使用实际的出生季度还是假的出生季度作为工具变量,这都不要紧,基本上,TSLS 会给出一样的答案!

这就是令劳动经济计量学家恐慌的一个回归。使用真实数据所计算的 TSLS 标准误,表明了教育的回报被精确地估计了,但是使用假数据所计算的标准误也是如此。当然,假数据不可能精确地估计出教育的回报,因为假的工具变量是完全不相关的。于是现在担心的是,基于实际数据的 TSLS 估计值就和基于假数据的 TSLS 估计值一样是不可靠的。

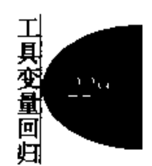
问题在于,Angrist 与 Krueger 的一些回归中的工具变量实际上是很弱的。在他们的一些设定中,第一阶段 F 统计量小于 2,远小于经验规则中的切割点 10。在其他的设定中,Angrist 与 Krueger 第一阶段的 F 统计量值较大,在这些设定中,TSLS 推断不受弱工具变量问题的影响。顺便提一下,在这些设定中,估计出的教育回报大约是 8%,稍大于用 OLS 所估计的值。①

注:①在 Angrist 与 Krueger(1991)中报道了最初的 IV 回归,而使用假工具变量的重新分析发表在 Bound,Jaeger 与 Baker(1995)中。

10.3.2 假设 2:工具变量外生性

如果工具变量不是外生的,那么 TSLS 就是不一致的:TSLS 依概率收敛于某个不同于回归中总体系数的值。毕竟,工具变量回归的思想就是工具变量包含了与误差项 u_i 无关的 X_i 变化的信息。事实上,如果工具变量不是外生的,那么它就不能正确指出 X_i 中的这个外生变化,而且有充分的理由认为 IV 回归没有提供一致性的估计量。附录 10.4 中总结了这个观点背后的数学原理。

能否在统计上检验"工具变量是外生的"这个假设？不能。更精确地说,假设你有和内生回归因子一样多的工具变量(系数被恰好识别),那么建立"这些工具变量实际上是外生的"这一假设的统计检验是不可能的,也就是说,经验证据不能被用来回答这些工具变量是否满足外生性限制条件这个问题。在这种情况下,评价所用的工具变量是否是外生的惟一方法是利用专家的观点和你对手中这一经验问题的个人知识。例如,Wright 父子关于农业



供给和需求方面的知识会使他们建议,低于平均的降雨量似乎会合理地使黄油供给曲线发生移动,但不会直接地使需求曲线发生移动。

评价一个工具变量是否是外生的,必须根据个人在某一应用方面的知识做出专家判断。然而,如果工具变量比内生回归因子多,那么存在一个有用的统计工具应用于这个过程,即所谓的过度识别约束检验。

过度识别约束检验。假设你有一个内生回归因子、两个工具变量,但没有内含外生变量,那么你可以计算两个不同的 TSLS 估计量,一个使用第一个工具变量,另一个使用第二个工具变量。由于抽样的变化,这两个估计量将不会是一样的。但是,如果这两个工具变量都是外生的,那么这两个估计量将趋于相互接近。但是如果这两个工具变量生成的估计量很不相同,那该怎么办呢? 此时你可能会明智地断言,这两个工具变量或者其中一个有问题,或者两个都有问题。也就是说,这两个工具变量或者其中一个,或者两个都不是外生的,得出这一结论应该是合理的。

过度识别约束的检验(test of overidentifying restrictions)以隐含的方式对这一问题做了比较。我们之所以说是以隐含的方式,是因为该检验是在实际上没有计算所有不同可能的 IV 估计值的情况下进行的。下面是检验的思想。工具变量的外生性意味着它们与 u_i 不相关。这表明工具变量应该与 $\hat{u}_i^{TSLS}$ 近似不相关,其中 $\hat{u}_i^{TSLS}=Y_0-(\hat{\beta}_0^{TSLS}+\hat{\beta}_1^{TSLS}X_{1i}+\cdots+\hat{\beta}_{k+r}^{TSLS}W_{ri})$ 是使用所有的工具变量所估计的 TSLS 回归的残差(是近似地而不是精确地,是因为存在抽样变差。注意,这些残差是用实际的 X 值而不是它们第一阶段的预测值来构造的)。因此,如果这些工具变量确实是外生的,那么在 $\hat{u}_i^{TSLS}$ 对工具变量和内含外生变量的回归中,这些工具变量的系数应该都是 0,这个假设是可以检验的。

计算过度识别约束检验的这一方法,在重要概念 10.6 中做了总结。检验的统计量是用仅适用于同方差的 F 统计量计算的。这个检验统计量通常被称为 J 统计量。

在大样本条件下,如果工具变量不是弱工具变量,而且误差是同方差的,那么在工具变量是外生的这一零假设下,这个 J 统计量具有服从自由度为 $m-k$ 的卡方分布(χ^2_{m-k})。需要记住的是,即使被检验约束条件的个数是 m,J 统计量的渐近分布的自由度也应该是 $m-k$,原因是只有可能检验 $m-k$ 个过度识别约束条件。

当系数被恰好识别($m=k$)时,为了弄清楚为什么不能检验回归因子的外生性,最容易的方法就是考虑只存在单个内含内生变量($k=1$)的情形。如果有两个工具变量,那么你可以计算两个 TSLS 估计量,各对应着一个工具变量,并且你可以比较它们,看它们是否接近。但是如果你只有一个工具变量,那么你只能计算一个 TSLS 估计量,你没有东西和它进行比较。事实上,如果系数被恰好识别,有 $m=k$,那么过度识别检验的 J 统计量正好是 0。

重要概念 10.6

过度识别约束检验(J 统计量)

假设 $\hat{u}_i^{TSLS}$ 是方程(10.12)中的 TSLS 估计的残差。使用 OLS 来估计下列方程中的回归系数:

$$\hat{u}_i^{TSLS}=\delta_0+\delta_1Z_{1i}+\cdots+\delta_mZ_{mi}+\delta_{m+1}W_{1i}+\cdots+\delta_{m+r}W_{ri}+e_i \tag{10.17}$$

其中,e_i 是回归误差项。假定 F 代表检验假设 $\delta_1=\cdots=\delta_m=0$ 的仅适用于同方差的 F 统计量。过度识别约束检验统计量是 $J=mF$。在所有工具变量都是外生的这一零假设下,在大样本条件下,J 服从 χ^2_{m-k} 分布,其中 $m-k$ 是"过度识别度",即工具变量的个数减去内生回归因子的个数。



10.4 在香烟需求案例中的应用[1]

前面我们试图用 TSLS 法估计香烟的需求弹性(见公式(10.16)),在公式(10.16)中,收入是个内含的外生变量,并且有两个工具变量,即总销售税和香烟特产税。现在我们放下这种方法,转到对这些工具变量的详细评价上。

和 10.1 节中说明的一样,这两个工具变量是相关的,因为税是香烟价格的一个大的组成部分,我们会马上从实证上对此进行分析。不过,我们首先关注这样一个难题,即这两个税收变量是否是外生变量。

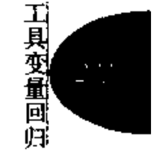
评价一个工具变量是否是外生变量,第一步是仔细考虑它为什么可能是或者为什么可能不是的争论。这要求我们考虑,是什么因素解释了香烟需求方程中的误差项,以及这些因素是否与工具变量相关。

为什么某些州的人均香烟消费量比其他州高呢?一个原因可能是各州之间的收入差异,但是州的收入包含在公式(10.16)中了,因此它不是误差项的一部分。另一个原因是,存在影响需求的历史因素。例如,种植烟草的州比其他大多数州具有更高的吸烟率。这个因素会与税收有关吗?很可能有关。如果烟草种植和香烟生产是一个州的重要产业,那么这些产业可以施加影响以维持一个低的香烟特产税。这表明,在香烟需求中有一个遗漏因素——该州是否种植烟草和生产香烟——可能与香烟特产税相关。

解决误差项与工具变量之间的这种可能相关关系的一个方法是,把该州烟草种植和香烟生产的规模信息包含进来,这就是我们在将收入作为回归因子包含在需求方程中时所采取的方法。但是由于我们有香烟消费的面板数据,因此存在另一种不同的方法,该方法不要求这个信息。正如第 8 章中所讨论的,面板数据能够消除那些在实体(州)间变化但不随时间发生变化的变量的影响,例如决定一个州烟草和香烟业规模大小的气候因素和历史环境。在第 8 章中给出了解决这个问题的两种方法:构造两个不同时期之间变量变化的数据,并使用固定效应回归。为了使这里的分析尽可能简单,我们采用前一种方法,并在两个不同年份之间变量变化的基础上采用 8.2 节中所介绍的那种回归。

两个不同年份之间的时间跨度长短,影响我们对估计弹性的解释。由于香烟是一种能使人上瘾的东西,因此香烟的价格变化改变消费的行为需花费一定的时间。起初,香烟价格上升可能对需求几乎没有影响。不过,过一段时间后,价格上升有助于强化一些吸烟者戒烟的愿望,而且重要的是,涨价可以打击那些非吸烟者进入吸烟者的行列。因此,在短期内,需求对价格上升的反应可能很小,但在长期内会很大。换句话说,对像香烟这样容易使人上瘾的产品,在短期内需求可能是缺乏弹性的,也就是说,它可能有个接近于 0 的短期弹性,但是从长期看它可能是更富有弹性的。

在这个分析中,我们集中于估计长期的价格弹性,通过考虑 10 年间数量和价格所发生的变化来估计弹性。具体来说,在此所考虑的回归方程中,10 年的对数数量变化 $\ln(Q_{i,1995}^{cigarettes})-\ln(Q_{i,1985}^{cigarettes})$ 对 10 年的对数价格变化 $\ln(P_{i,1995}^{cigarettes})-\ln(P_{i,1985}^{cigarettes})$ 以及 10 年的对数收入变化 $\ln(Inc_{i,1995})-\ln(Inc_{i,1985})$ 进行回归。使用的两个工具变量是:10 年间销售税的变化 $SalesTax_{i,1995}-SalesTax_{i,1985}$ 和 10 年间香烟特产税的变化 $CigTax_{i,1995}-CigTax_{i,1985}$。



① 本节假设你了解 8.1 节和 8.2 节中 $T=2$ 期的面板数据的资料。

一般兴趣框

吸烟的外部性

吸烟所强加的成本不单单是只由吸烟者本人完全承担,也就是说,吸烟产生了外部性,因此,征收香烟税的一个经济理由就是将这些外部性"内部化"。理论上,每包香烟的税收应该等于吸每包香烟所产生的外部性的美元价值,但是,以美元测度的吸每包香烟所产生的外部性确切地说是多少呢?

已有的几项研究使用了经济计量方法估计吸烟的外部性。由其他人所承担的负外部性——成本——包括政府为了照顾有病的吸烟者所支付的医疗费用,因间接吸烟所引致的非吸烟者的保健费用以及由香烟所引起的火灾损失。

但是,从纯粹的经济学观点来看,吸烟也有正的外部性或利益。吸烟的最大好处就是吸烟者在社会保障(公共养老金)税方面倾向于支付的比他们曾得到的回报多得多。在吸烟老人的家庭护理费用上也有大量的节省,因为吸烟者倾向于不会那么长寿。由于在吸烟者活着的时候吸烟产生负的外部性,但在其死后产生正外部性的累积,因此,每包香烟外部性的净现值(每包香烟的净成本值贴现到现在)依赖于贴现率。

该研究并没有在净外部性的具体美元价值上取得一致意见。一些研究指出,经过适当贴现以后的净外部性是很小的,小于当前的税收额。实际上,最极端的估计表明,净外部性是正的,因此吸烟应该受到补贴。另一些研究表明(这些研究考虑了那些可能是重要的但却很难量化的成本,例如照顾那些因母亲吸烟而不健康的婴儿),每包香烟的外部性可能是1美元,甚至更多,但是所有的研究都一致认为,中年后期死亡的吸烟者所支付的税比他们在短暂的退休期间所得到的回报要多。①

注:①关于吸烟外部性的早期计算结果是由 Willard G. Manning 等(1989)报告的。Barendregt 等(1997)所给出的计算表明,如果每个人都戒烟,那么保健成本会上升。关于吸烟外部性的其他研究成果,由 Chaloupka 和 Warner(2000)做了综述。

表10—1　　使用美国48个州的面板数据,香烟需求的两阶段最小二乘估计值

因变量:$\ln(Q_{i,1995}^{cigarettes})-\ln(Q_{i,1985}^{cigarettes})$			
回归因子	(1)	(2)	(3)
$\ln(P_{i,1995}^{cigarettes})-\ln(P_{i,1985}^{cigarettes})$	-0.94**	-1.34**	-1.20**
	(0.21)	(0.23)	(0.20)
$\ln(Inc_{i,1995})-\ln(Inc_{i,1985})$	0.53	0.43	0.46
	(0.34)	(0.30)	(0.31)
截距	0.21	0.45**	0.37**
	(0.13)	(0.14)	(0.12)
工具变量	销售税	香烟特产税	销售税和香烟特产税
第一阶段 F 统计量	33.70	107.20	88.60
过度识别约束 J 检验和	—	—	4.93
P 值			(0.026)

注:这些回归是利用美国48个州的数据(10年差分的48个观测值)来估计的。附录10.1描述了这个数据。在重要概念10.6中描述了过度识别约束的 J 检验(它的 p 值在括号中给出),而在重要概念10.5中描述了第一阶段 F 统计量。单个系数在*5%的显著性水平下或**1%的显著性水平下在统计上是显著的。



结果在表 10—1 中给出。和通常一样，表中每一列表示不同回归的结果。所有回归都用相同的回归因子，并都使用 TSLS 估计所有的系数，这三个回归之间的惟一差别就是所使用的工具变量组不同。在第(1)列中，惟一的工具变量是销售税；在第(2)列中，惟一的工具变量是香烟特产税；在第(3)列中，两个税收都被用做工具变量。

在 IV 回归中，系数估计值的可靠性取决于工具变量的有效性，因此，在表 10—1 中，所要考虑的第一个问题就是评价工具变量有效性的诊断统计量。

第一，工具变量是相关的吗？三个回归中的第一阶段 F 统计量分别是 33.7，107.2 和 88.6，在所有这三个回归中，第一阶段 F 统计量都超过了 10。我们由此断言，这些工具变量都不是弱工具变量，因此，我们可以依赖所估计的系数和标准误的统计推断（假设检验和置信区间）的标准方法。

第二，工具变量是外生的吗？由于第(1)列和第(2)列中的每个回归都含有一个工具变量和一个内含内生回归因子，因此这些回归中的系数都被恰好识别。这样，在这两个回归中，我们就不能进行 J 检验了。然而，第(3)列中的回归被过度识别，因为有两个工具变量和一个内含回归因子，所以有一个($m-k=2-1=1$)过度识别约束。J 统计量是 4.93，它服从χ_1^2 分布，因此，5%的临界值是 3.84（见附表 3），两个工具变量都是外生的这一零假设在 5%的显著性水平下被拒绝（这个推断也可以根据表 10—1 中所给出的 0.026 的 p 值直接得出）。

J 统计量拒绝的原因是，这两个工具变量生成了根本不同的估计系数。当惟一的工具变量是销售税（第(1)列）时，估计的价格弹性是 -0.94，但是当惟一的工具变量是香烟特产税时，所估计的价格弹性是 -1.34。回忆一下 J 统计量的基本思想：如果两个工具变量都是外生的，那么使用单个工具变量的两个 TSLS 估计量是一致的，其值只会因随机抽样变差而互不相同。然而，如果一个工具变量是外生的，而另一个不是，那么基于内生工具变量的估计量就是不一致的。J 统计量就可洞悉这一点。在这个应用中，这两个所估计的价格弹性之间的差异相当大，因此它不可能是纯粹抽样变差的结果，因此 J 统计量拒绝了两个工具变量都是外生的这个零假设。

J 统计量拒绝意味着第(3)列中的回归是以无效的工具变量为基础的（工具变量外生性条件不成立），那第(1)列和第(2)列中的估计值意味着什么呢？J 统计量拒绝说明，至少有一个工具变量是内生的，因此有三种逻辑可能性：销售税是外生的，但香烟特产税不是，在此情形下，第(1)列中的回归是可靠的；香烟特产税是外生的，但销售税不是，因此，第(2)列回归是可靠的；两个税都不是外生的，因此两个回归都不可靠。统计证据并不能告诉我们哪一种可能性是正确的，所以我们必须进行判断。

我们认为总销售税的外生性要比香烟特产税的外生性强，因为政治过程能够将香烟特产税变化和香烟市场以及吸烟政策变化联系起来。例如，如果一个州的吸烟量因吸烟变得过时而降低，那么将会有更少的吸烟者，反对香烟特产税增加的游说声音也会减弱，这反过来会导致更高的香烟特产税。这样，偏好（这是 u 的一部分）的变化可能与香烟特产税（工具变量）的变化相关。这表明，要对将惟一的香烟税作为工具变量所得出的 IV 估计值打折扣。这里相应的建议是，只采用将总销售税作为工具变量所估计的价格弹性值 -0.94。

-0.94 的估计值表明了香烟消费并不是非常缺乏弹性的：价格增加 1% 导致香烟消费减少 0.94%。对香烟这样使人上瘾的产品，这看上去可能是令人惊讶的，但是要记住，这个弹性是用 10 年间的数据变化计算的，所以它是长期弹性。这个估计值表明，增税能够对香烟消费产生巨大影响，至少在长期是这样。



这两个效应的一些复杂的组合。这个问题不可能通过找到更好的控制变量来解决。

不过,这个联立因果偏差可以通过找到一个合适的工具变量,使用 TSLS 来消除。找到的工具变量必须与监禁率相关(即它必须是相关的),但是它也一定与我们所研究的犯罪率方程中的误差项不相关(即它必须是外生的)。也就是说,它必须影响监禁率,但是与决定犯罪率的任何未观测到的因素又不相关。

到哪里去找那些虽影响监禁率但对犯罪率没有直接影响的因素呢?一处就是反映现有监狱容量外生变化的因素。由于建造监狱花费的时间长,因此,短期容量约束会迫使州政府提前释放囚犯,或者减少监禁率。使用这种推理,Levitt(1996)建议可将针对减少监狱过度拥挤的诉讼作为一个工具变量,并且他使用从 1972 年到 1993 年美国州的面板数据证实了这个思路。

测度过度拥挤诉讼的变量是有效的工具变量吗?虽然在 Levitt 的研究数据中他没有给出第一阶段 F 统计量,但是监狱过度拥挤的诉讼减缓了囚犯监禁的增长,这表明这个工具变量是相关的。由于过度拥挤诉讼是由监狱条件引起的,而不是由犯罪率或它的决定性因素引起的,因此,从这个意义上说,这个工具变量是外生的。由于 Levitt 将过度拥挤诉讼分解成好几个类型,这样就有好几个工具变量,因此他能够检验过度识别约束,使用 J 检验不能拒绝它们,这就支持了他的工具变量是有效的这一结论。

Levitt 使用这些工具变量和 TSLS 估计了监禁对犯罪率的效应是巨大的。这个估计效应比用 OLS 估计的效应大 3 倍,这表明 OLS 遭受了很大的联立因果偏差。

削减班级规模会提高考试成绩吗?如我们在第 2 部分的实证分析中所看到的,具有小班的学校倾向于更富有,学生在学校内外拥有更多的学习机会。在第 2 部分中,我们通过控制学生富裕程度、说英语的能力等的不同测度,使用多元回归方法处理遗漏变量偏差的威胁。怀疑者仍然想知道我们做的是否充分,如果我们遗漏了重要的东西,我们的班级规模效应的估计值仍然会有偏差。

这个潜在的被遗漏变量的偏差可以通过引入恰当的控制变量来解决,但是如果这些数据是不可获得的(有些难于测度,像校外学习机会),那么一个可供选择的方法是使用 IV 回归。这个回归要求存在一个与班级规模相关(相关性),但是与构成误差项的被遗漏的考试成绩的决定性因素(例如,家长的学习兴趣、校外的学习机会、教师的质量和学校设备等等)不相关(外生性)的工具变量。

到哪去寻找这个引起班级规模发生随机性外生变差,但与考试成绩的其他决定性因素又无关的工具变量呢?Hoxby(2000)建议在生物学领域寻找。由于出生日期随机波动,因此新入幼儿园的班级规模在不同年份间是变化的。虽然进入幼儿园的实际儿童数可能是内生的(关于学校的近期新闻可能会影响家长是否将孩子送到私立学校),但是她论证说进入幼儿园的潜在儿童数量——该地区内的 4 岁儿童数——主要是儿童出生日期随机波动的问题。

潜在入学人数是个有效的工具变量吗?它是否是外生的依赖于它是否与未观测到的班级规模的决定性因素相关。潜在入学人数在生物学上的波动确实是外生的,但是潜在入学人数也会因幼童家长选择搬进正在改进的校区和搬出处于困境中的校区而波动。如果是这样的话,那么潜在入学人数的增加可能与诸如学校管理质量这样难以观测的因素相关,进而使这个工具变量变得无效。Hoxby 通过以下的推理来解决这个问题:潜在学生群体的增长或下降会在几年内平稳地发生,而出生日期的随机波动会在潜在的入学人数上产生短期的"尖峰"。因此,她不将潜在入学人数,而是将潜在入学人数与它的长期趋势的离差作为工



具变量。这些离差满足工具变量的相关性准则(第一阶段 F 统计量都超过了100)。她举了一个很好的例子说明该工具变量是外生的,但是和在所有IV分析中一样,这个假设的可靠性最终还是一个判断性问题。

Hoxby使用康涅狄格州在20世纪80年代和90年代小学的详细面板数据,证实了她的这个策略。该面板数据集允许她把学校固定效应包括进来,除了工具变量策略之外,学校固定效应解决了在学校层次出现的遗漏变量偏差问题。她的TSLS估计值表明,班级规模对考试成绩的效应是很小的,她的大部分估计值在统计上都是不显著地异于0的。

对心脏病人进行积极治疗会延长寿命吗?对心脏病(学术上称为急性心肌梗塞或AMI)患者而言,新的积极治疗挽救了潜在病人的生命。在一种新的疗法——在这个例子中是心肌导管插入术[①]——被批准普遍使用之前,它要经历临床实验,即设计一系列的随机化控制实验来测度它的作用和副作用,但是,临床试验中效果好是一回事,现实中的实际效果则是另一回事。

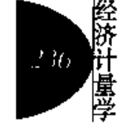
估计心肌导管插入术在现实中效果的一个自然的出发点是,比较接受这种治疗的病人和没有接受这种治疗的病人,这会导致用病人存活的时间对二元处理变量(病人是否接受心肌导管插入术)和其他影响死亡的控制变量(年龄、体重、其他可测度的健康条件等等)进行回归。该指示变量的总体系数就是这种疗法所提供的病人预期寿命的增量。不幸的是,OLS估计量受偏差的影响:心肌导管插入术并不是随机地"正好发生"在一个病人身上,更确切地说,它是因为病人和医生断定它可能会有效才被执行的。如果他们(她们)的决策部分地以和健康结果有关的不在数据集中的那些难以观测的因素为基础,那么治疗决策将会与回归误差项相关。如果最健康的病人是接受该项治疗的病人,那么OLS估计量将是有偏的(治疗与遗漏变量相关),则该项治疗看上去将比它实际的情况更有效。

使用一个有效的工具变量,通过IV回归,可以消除这个潜在的偏差。该工具变量必须与治疗相关(即必须是相关的),但同时又必须与影响生存的被省略的健康因素不相关(即必须是外生的)。

除了通过它的治疗效果外,到哪里去寻找影响治疗但又不影响健康结果的因素呢? McClellan, McNeil与Newhouse(1994)建议从地理学角度去寻找。他们数据集中的大部分医院并不专攻心肌导管插入术.因此,许多病人更接近不提供这种治疗的"常规"医院,而不接近提供心肌导管插入术的医院。所以,McClellan, McNeil与Newhouse将AMI病人的家到最近的心肌导管插入术医院的距离和到最近的任何类型医院的距离之差作为工具变量。如果最近的医院是心肌导管插入术医院,那么这个距离就是0;否则,它就是正的。如果这个相对距离影响了接受这种治疗的概率,那么它是相关的。如果在AMI患者间它是随机分布的,那么它是外生的。

到最近的心肌导管插入术医院的相对距离是个有效的工具变量吗? McClellan, McNeil与Newhouse没有给出第一阶段的 F 统计量,但是他们却提供了它不是弱工具变量的其他经验证据。这个距离测度是外生的吗? 他们给出了两点理由。第一,他们利用他们的医学专业知识和保健系统知识,认为到医院的距离与任何难以观测的决定AMI结果的变量似乎是不相关的。第二,他们拥有一些影响AMI结果的其他变量的数据,如病人的体重,并且在他们的样本中,距离与这些可观测的生存决定性因素是不相关的。他们认为,这使距离与误差项中难以观测到的决定性因素之间不相关更为可信。

① 心肌导管插入术是一种疗法,在这种疗法中,导管或者管子被插入到血管中,并被引向到达心脏的所有路径以获得关于心脏和冠状动脉的信息。



总结

1. 工具变量回归是当一个或多个回归因子与误差项相关时估计回归系数的一种方法。

2. 在所研究的回归方程中，内生变量与误差项相关，外生变量与该误差项不相关。

3. 要使工具变量是有效的，则：(1)它一定与内含内生变量相关；(2)它一定是外生的。

4. IV 回归要求至少有与内含内生变量一样多的工具变量。

5. TSLS 估计量有两个阶段：第一阶段，内含内生变量对内含外生变量和工具变量进行回归；第二阶段，因变量对内含外生变量和由第一阶段回归中得出的内含内生变量的预测值进行回归。

6. 弱工具变量(与内含内生变量几乎不相关的工具变量)使 TSLS 估计量有偏，而且使 TSLS 置信区间和假设检验不可靠。

7. 如果一个工具变量不是外生的，那么其 TSLS 估计量是不一致的。

重要术语

工具变量(IV)回归　工具变量(工具)　内生变量　外生变量　工具变量相关性条件　工具变量外生性条件　两阶段最小二乘　内含外生变量　恰好识别　过度识别　不足识别　简约式　第一阶段回归　第二阶段回归　弱工具变量　第一阶段 F 统计量　过度识别约束检验

复习概念

10.1　在方程(10.3)的需求曲线回归模型中，$\ln(P_i^{butter})$ 与误差 u_i 是正相关还是负相关？如果用 OLS 估计 β_1，你认为 β_1 的估计值比它的真值大还是小呢？请解释说明。

10.2　在本章香烟需求的案例研究中，假设我们将这个州的人均树木数量作为一个工具变量，这个工具变量是相关的吗？它是外生的吗？它是个有效的工具变量吗？

10.3　在监禁对犯罪率影响的研究中，假设 Levitt 将人均律师数作为工具变量，这个工具变量是相关的吗？它是外生的吗？它是个有效的工具变量吗？

10.4　在心肌导管插入术效果的研究中，McClellan，McNeil 与 Newhouse(1994)将病人到心肌导管插入术医院的距离和到常规医院的距离之差作为工具变量。你如何决定这个工具变量是否是相关的？你如何决定这个工具变量是否是外生的？

练习

*10.1　这个问题涉及表 10—1 中所总结的面板数据回归。



a. 假设联邦政府正考虑一种关于香烟的新税收，估计会使每包香烟的零售价格增加 0.10 美元，如果每包香烟的当前价格是 2.00 美元，使用第(1)列中的回归预测需求的变化，构造需求变化的 95% 的置信区间。

b. 假设美国进入经济衰退期，收入下降 2%，使用第(1)列中的回归预测需求的变化。

c. 通常衰退期持续不到一年。你认为第(1)列中的回归会为问题(b)提供一个可靠的

答案吗？为什么？

d. 假设第(1)列中的 F 统计量是3.6，而不是33.6，那么这个回归会为(a)中所提出的问题提供一个可靠的答案吗？为什么？

10.2　考虑含有一个回归因子的回归模型：$Y_i=\beta_0+\beta_1X_i+u_i$，假定满足重要概念4.3中的假设。

a. 证明：X_i 是个有效的工具变量，也就是说，证明重要概念10.3满足 $Z_i=X_i$。

b. 证明：重要概念10.4中的IV回归假设满足 Z_i 的选择。

c. 证明：使用 $Z_i=X_i$ 所构造的IV估计量等同于OLS估计量。

10.3　一个同学对估计方程(10.1)中误差项的方差感兴趣。

a. 假设她使用TSLS第二阶段回归中的估计量：$\hat{\sigma}_a^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{TSLS}-\hat{\beta}_1^{TSLS}\hat{X}_i)^2$，其中 $\hat{X}_i$ 是来自于第一阶段回归的拟合值。这个估计量是一致的吗？（为了回答这个问题，假设样本量非常大，TSLS估计量实质上等同于 β_0 和 β_1）

b. $\hat{\sigma}_b^2=\frac{1}{n-2}\sum_{i=1}^{n}(Y_i-\hat{\beta}_0^{TSLS}-\hat{\beta}_1^{TSLS}X_i)^2$ 是一致的吗？

10.4　考虑一个含有单个内含内生变量和单个工具变量的TSLS估计，那么，第一阶段回归预测值是 $\hat{X}_i=\hat{\pi}_0+\hat{\pi}_1Z_i$。使用样本方差和样本协方差的定义证明 $s_{\hat{X}Y}=\hat{\pi}_1s_{ZY}$ 和 $s_{\hat{X}}^2=\hat{\pi}_1^2s_Z^2$。使用这个结论填写附录10.2中公式(10.4)的推导步骤。

附录10.1　香烟消费面板数据集

该数据集由1985年到1995年美国48个大陆州的年度数据构成。香烟消费量是用每财政年度里以包为单位计算的人均香烟销售量测度的，它是从州税收数据中导出的。价格是该财政年度期间每包香烟的平均零售价，包括税收。收入是人均收入。由于广泛基准的州销售税适用于所有的消费品，因此一般销售税就是以美分计算的每包香烟的平均税。香烟特产税是只适用于香烟的税。本章回归分析中所使用的所有价格、收入和税收均用消费者价格指数进行了缩减，因此它们是以实际美元即不变价计算的。我们感谢MIT的Jonathan Gruber教授向我们提供了这些数据。

附录10.2　公式(10.4)中TSLS估计量公式的推导

TSLS的第一阶段是用 X_i 对工具变量 Z_i 进行OLS回归，并计算OLS预测值 $\hat{X}_i$，而第二阶段就是用 Y_i 对 $\hat{X}_i$ 进行OLS回归。因此，TSLS估计量的公式(用预测值 $\hat{X}_i$ 来表达)，就是重要概念4.2中OLS估计量的公式，其中用 $\hat{X}_i$ 代替了 X_i，即 $\hat{\beta}_1^{TSLS}=s_{\hat{X}Y}/s_{\hat{X}}^2$，其中 $s_{\hat{X}}^2$ 是 $\hat{X}_i$ 的样本方差，$s_{\hat{X}Y}$ 是 Y_i 和 $\hat{X}_i$ 之间的样本协方差。

由于 $\hat{X}_i$ 是 X_i 在第一阶段回归中的预测值，$\hat{X}_i=\hat{\pi}_0+\hat{\pi}_1Z_i$，所以样本方差和协方差的定义暗含着 $s_{\hat{X}Y}=\hat{\pi}_1s_{ZY}$ 和 $s_{\hat{X}}^2=\hat{\pi}_1^2s_Z^2$（见练习10.4）。因此，TSLS估计量可被写成 $\hat{\beta}_1^{TSLS}=s_{\hat{X}Y}/s_{\hat{X}}^2=s_{ZY}/(\hat{\pi}_1s_Z^2)$。最后，$\hat{\pi}_1$ 是来自TSLS第一阶段的OLS斜率系数，所以 $\hat{\pi}_1=s_{ZX}/s_Z^2$。将这个表达式中的 $\hat{\pi}_1$ 代入公式 $\hat{\beta}_1^{TSLS}=s_{ZY}/(\hat{\pi}_1s_Z^2)$，便得到了公式(10.4)中的TSLS估计量公式。



附录 10.3 TSLS 估计量的大样本分布

本附录研究了在 10.1 节中所考虑的条件下，即有单个工具变量、单个内含内生变量，而没有内含外生变量的条件下，TSLS 估计量的大样本分布。

一开始，我们以误差的形式导出了 TSLS 估计量的一个公式，这个公式构成了下面讨论的基础，类似于附录 4.3 中公式(4.51)的 OLS 估计量的表达式。根据公式(10.1)有：$Y_i-\overline{Y}=\beta_1(X_i-\overline{X})+(u_i-\bar{u})$，因此，$Z$ 和 Y 之间的样本协方差可被表示为：

$$\begin{aligned}s_{ZY}&=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})(Y_i-\overline{Y})\\&=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})[\beta_1(X_i-\overline{X})+(u_i-\bar{u})]\\&=\beta_1s_{ZX}+\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})(u_i-\bar{u})\\&=\beta_1s_{ZX}+\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})u_i\end{aligned}\tag{10.18}$$

其中，$s_{ZX}=\frac{1}{n-1}\sum_{i=1}^{n}(Z_i-\overline{Z})(X_i-\overline{X})$，而且这里最后一个等式是由 $\sum_{i=1}^{n}(Z_i-\overline{Z})=0$ 推导出来的。将 s_{ZY}的定义和方程(10.18)中最后的表达式代入 $\hat{\beta}_1^{TSLS}$ 的定义中，并用$(n-1)/n$ 同时乘以分式的分子和分母，得到：

$$\hat{\beta}_1^{TSLS}=\beta_1+\frac{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})u_i}{\frac{1}{n}\sum_{i=1}^{n}(Z_i-\overline{Z})(X_i-\overline{X})}\tag{10.19}$$

当重要概念 10.4 中的 IV 回归假设成立时，$\hat{\beta}_1^{TSLS}$ 的大样本分布

TSLS 估计量的公式(10.19)与附录 4.3 中的 OLS 估计量的公式(4.51)类似，只不过分子中出现的是 Z 而不是 X，分母中是 Z 和 X 的协方差而不是 X 的方差。由于这些相似性，以及 Z 是外生的，因此附录 4.2 中关于 OLS 估计量在大样本条件下服从正态分布的论点可推广到$\hat{\beta}_1^{TSLS}$。

具体来说，当样本量很大时，$\overline{Z}\approx\mu_Z$，因此分子大约是 $\bar{q}=\frac{1}{n}\sum_{i=1}^{n}q_i$，其中 $q_i=(Z_i-\mu_Z)u_i$。因为工具变量是外生的，所以 $E(q_i)=0$。根据重要概念 10.4 中的 IV 回归假设，q_i 是独立同分布的，其方差为 $\sigma_q^2=\text{var}[(Z_i-\mu_Z)u_i]$。由此得出 $\text{var}(\bar{q})=\sigma_{\bar{q}}^2=\sigma_q^2/n$，而根据中心极限定理，在大样本条件下，$\bar{q}/\sigma_{\bar{q}}$服从 $N(0,1)$分布。

由于样本协方差对总体协方差是一致的，$s_{ZX}\xrightarrow{P}\text{cov}(Z_i,X_i)$，这里的工具变量是相关的，因此 $\text{cov}(Z_i,X_i)$是非零的。所以，根据公式(10.19)，$\hat{\beta}_1^{TSLS}\approx\beta_1+\bar{q}/\text{cov}(Z_i,X_i)$，因此在大样本条件下，$\hat{\beta}_1^{TSLS}$ 近似服从分布 $N(\beta_1,\sigma^2_{\hat{\beta}_1^{TSLS}})$，其中，$\sigma^2_{\hat{\beta}_1^{TSLS}}=\sigma_{\bar{q}}^2/[\text{cov}(Z_i,X_i)]^2=(1/n)\text{var}[(Z_i-\mu_Z)u_i]/[\text{cov}(Z_i,X_i)]^2$，这就是公式(10.8)中所给出的表达式。

附录 10.4 当工具变量无效时，TSLS 估计量的大样本分布

本附录考虑当一个或另一个工具变量有效性的条件不成立时，在 10.1 节的结构(一个

X,一个 Z)下 TSLS 估计量的大样本分布。如果工具变量相关性条件不成立(即工具变量是弱工具变量),那么 TSLS 估计量的大样本分布不是正态分布,实际上,它的分布是两个正态随机变量之比的分布。如果该工具变量的外生性条件不成立,那么 TSLS 估计量是不一致的。

当工具变量是弱工具变量时,$\hat{\beta}_1^{TSLS}$ 的大样本分布

首先考虑工具变量为不相关的情形,有 $\text{cov}(Z_i,X_i)=0$。这样,附录 10.3 中的变量必须被 0 除。为了避免这个问题,当总体协方差为 0 时,我们需要进一步地考察等式(10.19)中分母的行为。

我们重写公式(10.19)。由于在大样本条件下样本均值的一致性,$\overline{Z}$ 接近于 μ_Z,$\overline{X}$ 接近于 μ_X,因此,公式(10.19)中的分母项近似为 $\frac{1}{n}\sum_{i=1}^{n}(Z_i-\mu_Z)(X_i-\mu_Z)=\frac{1}{n}\sum_{i=1}^{n}r_i=\bar{r}$,其中 $r_i=(Z_i-\mu_Z)(X_i-\mu_X)$。设 $\sigma_r^2=\text{var}[(Z_i-\mu_Z)(X_i-\mu_X)]$,$\sigma_{\bar{r}}^2=\sigma_r^2/n$,并且设 $\bar{q}$,σ_q^2 和 $\sigma_{\bar{q}}^2$ 为附录 10.3 中所定义的内容,那么在大样本条件下,方程(10.19)隐含着:

$$\hat{\beta}_1^{TSLS}\approx\beta_1+\frac{\bar{q}}{\bar{r}}=\beta_1+\left(\frac{\sigma_{\bar{q}}}{\sigma_{\bar{r}}}\right)\left(\frac{\bar{q}/\sigma_{\bar{q}}}{\bar{r}/\sigma_{\bar{r}}}\right)=\beta_1+\left(\frac{\sigma_q}{\sigma_r}\right)\left(\frac{\bar{q}/\sigma_{\bar{q}}}{\bar{r}/\sigma_{\bar{r}}}\right) \tag{10.20}$$

如果该工具变量是不相关的,那么 $E(r_i)=\text{cov}(Z_i,X_i)=0$。这样,$\bar{r}$ 就是随机变量 r_i 的样本均值(根据第二阶段最小二乘假设),其中随机变量 r_i,$i=1,\cdots,n$,是独立同分布的,有方差 $\sigma_r^2=\text{var}[(Z_i-\mu_Z)(X_i-\mu_X)]$(根据 IV 回归第三个假设它是有限的)且有零均值(由于工具变量是不相关的)。由此得出结论,中心极限定理适用于 $\bar{r}$,具体地讲,$\bar{r}/\sigma_{\bar{r}}$ 渐近服从分布 $N(0,1)$。因此,公式(10.20)的最后表达式意味着在大样本条件下,$\hat{\beta}_1^{TSLS}-\beta_1$ 的分布是的 aS 分布,其中 $a=\sigma_q/\sigma_r$,S 是两个随机变量之比,其中每个随机变量都服从标准正态分布(这两个标准正态随机变量是相关的)。

换句话说,当这个工具变量不相关时,中心极限定理既适用于 TSLS 估计量的分母,也适用于其分子,因此在大样本条件下,TSLS 估计量的分布是两个正态随机变量之比的分布。因为 X_i 和 u_i 是相关的,所以这些正态随机变量是相关的,并且当工具变量不相关时,TSLS 估计量的大样本分布很复杂。实际上,具有不相关的工具变量的 TSLS 估计量的大样本分布是以 OLS 估计量的概率极限为中心的。因而,当工具变量不相关时,TSLS 并没有消除 OLS 中的偏差。此外,TSLS 估计量服从非正态分布,即使在大样本条件下。

当工具变量是弱的但并不是不相关的时,TSLS 估计量的分布仍然是非正态分布,因此,这里关于不相关工具变量极端情况的一般结论可延伸到弱工具变量上。例如,我们可以证明,在大样本条件下,TSLS 估计量抽样分布的均值约等于 $\beta_1+(\beta_1^{OLS}-\beta_1)/[E(F)-1]$,其中 β_1^{OLS} 是该 OLS 估计量的(概率)极限,即 $\hat{\beta}_1\xrightarrow{p}\beta_1^{OLS}$,$E(F)$ 是第一阶段 F 统计量的期望。这个关于 TSLS 估计量均值的表达式是重要概念 10.5 中所提出的弱工具变量经验规则诊断中切割值的来源。具体来说,如果 $E(F)=10$,那么相对于 OLS 的大样本偏差来说,TSLS 的大样本偏差是 1/9,或刚刚超过 10%,在许多应用中,这是人们可以接受的一个很小的数值。



当工具变量是内生的时,$\hat{\beta}_1^{TSLS}$ 的大样本分布

公式(10.19)最后表达式中的分子依概率收敛于 $\text{cov}(Z_i,u_i)$。如果该工具变量是外生的,那么 $\text{cov}(Z_i,u_i)$ 是 0,TSLS 估计量是一致的(假设工具变量不是弱工具变量)。然而,如

果工具变量不是外生的，那么在工具变量不是弱工具变量的情况下，$\hat{\beta}_1^{TSLS} \xrightarrow{p} \beta_1 + \text{cov}(Z_i, u_i)/\text{cov}(Z_i, X_i) \neq \beta_1$。也就是说，如果该工具变量不是外生的，那么 TSLS 估计量就是不一致的。

所讨论的,这些威胁中的一部分可用回归方法,包括“差分再差分”估计量和工具变量回归来解决或评估。11.4 节运用这些方法分析了 20 世纪 80 年代后期在田纳西州的一个随机化控制实验。在该实验中,小学生被随机地分配到不同规模的班级中去。

11.5 节转向准实验的讨论和如何使用准实验估计因果效应。对准实验有效性的威胁在 11.6 节中讨论。在实验和准实验中都可能出现的一个问题是,处理效应在同一总体中的不同成员之间可能不同。当总体是异质的时,对因果效应相应的估计值如何解释的问题,将在 11.7 节中讨论。

11.1 理想化实验和因果效应

回想一下 1.2 节中的内容,一项随机化控制实验将随机地从所研究的总体中选择主体(个体,或更一般地讲,实体),然后随机地将他们分配到接受实验处理的处理组中去,或不接受处理的控制组中去。该项处理的因果效应就是在理想的随机化控制实验中所测度的对所研究的处理产生结果的期望效应。

11.1.1 理想的随机化控制实验

起初,人们可能认为理想的实验就是取两个各个方面都相同的个体,处理他们中的一个,并在保持其他所有影响都不变的情况下比较他们结果的差异性。不过,这并不是一个实际的实验设计,因为不可能找到两个同样的个体。即使是完全相同的双胞胎,各自也会有不同的生活经历,因此他们不是在每个方面都相同的。

一个理想的随机化控制实验的中心思想就是,可以通过从一个总体中随机地选择个体,然后对一些个体随机地施行处理,进而测度因果效应。如果处理是随机分配的,例如,通过投掷硬币或通过使用计算机的随机数发生器,那么该处理的水平独立分布于对结果的任何其他决定性因素之中,由此便消除了遗漏变量偏差(见重要概念 5.1)的可能性。例如,假设个人被随机地分配参加一项工作培训项目,在该培训项目结束后,一个人以前的工作经验将影响他(或她)获得工作的机会,但是只要该工作培训项目(“处理”)中的参加者是随机分配的,那么工作经验的分布在处理组和控制组之间就是一样的,也就是说,参与工作培训是独立分布于以前的工作经验的。因此,参与同以前的工作经验是不相关的,在分析中忽略以前的工作经验,将不会使培训项目对将来就业效应的估计量产生遗漏变量偏差。

随机分配的作用可以用单个因子回归模型来重新表述:

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{11.1}$$

其中,X_i 是处理的水平,和通常一样,u_i 包含了结果 Y_i 的所有其他的决定性因素。如果处理对处理组的所有成员都是一样的,那么 X_i 就是二元变量。其中,$X_i=1$ 表示第 i 个个体接受了处理;$X_i=0$ 表示他(或她)没有接受处理。如果处理的水平在处理组的个体间有变化,那么 X_i 就是接受处理的水平。例如,X_i 可能是药的剂量或一项工作培训项目的周数,其中,如果没有接受处理(药的剂量为 0),那么 $X_i=0$。如果 X_i 是二元变量,那么公式(11.1)中的线性回归函数就不会施加函数形式约束。如果 X_i 能取多个值,那么公式(11.1)就会把总体回归函数看做是线性的(非线性问题可用 6.2 节的方法加以解决)。

如果 X_i 是随机分配的,那么 X_i 独立分布于 u_i 中的遗漏因素。由于这些遗漏因素与 X_i 是独立分布的,因此在公式(11.1)中,$E(Y_i|X_i)=\beta_0+\beta_1 X_i$。换句话说,给定 X_i 下,u_i 的条件均值不依赖于 X_i,即 $E(u_i|X_i)=0$。这样,X_i 的随机分配意味着,具有单个回归因子的回

归模型中的第一个最小二乘假设(见重要概念4.3)自动成立。

因果效应。处理水平 X 对 Y 的因果效应(casual effect)就是条件期望之差,即 $E(Y|X=x)-E(Y|X=0)$,其中 $E(Y|X=x)$ 是在一个理想的随机化实验中收到处理水平 x 的处理组中 Y 的期望值,而 $E(Y|X=0)$ 是控制组中 Y 的期望值。就实验来说,因果效应也被称为处理效应(treatment effect)。由于随机分配,在公式(11.1)中 $E(u_i|X_i)=0$,因此,公式(11.1)中的 β_1 就是用处理组和控制组之间结果的期望值之差所测度的 X 单位变化的因果效应。

11.1.2 差分估计量

因果效应是期望值之差,因而它是一个未知的总体特征。因果效应可以用取自随机化控制实验的数据进行估计。假设处理 X_i 是二元变量,由于这里的处理是随机分配的,因此因果效应可用处理组和控制组之间的样本平均结果之差来估计。同理,如在4.7节中所讨论的,β_1 可用 Y_i 对 X_i 回归的OLS估计量 $\hat{\beta}_1$ 进行估计。由于在公式(11.1)中 $E(u_i|X_i)=0$,因此 $\hat{\beta}_1$ 是无偏的。我们把 Y_i 对 X_i 回归的OLS估计量 $\hat{\beta}_1$ 称为差分估计量(differences estimator),当处理是二元变量时,它是处理组的样本平均结果和控制组的样本平均结果之差。



通过随机地分配处理,一个理想的随机化控制实验消除了处理 X_i 和误差项 u_i 之间的相关关系,所以差分估计量是无偏的且一致的。然而在实践中,现实的实验与理想的实验总是有偏离,于是就会出现 X_i 和 u_i 之间相关性的问题。

11.2 现实中的实验存在的潜在问题

回忆一下重要概念7.1,如果一项关于因果效应的统计推断对所研究的总体是有效的,那么就称该项统计研究是内部有效的;如果它的推断和结论能够从所研究的总体和环境中推广到其他的总体和环境中,那么该项统计研究就是外部有效的。各种现实世界中的问题向以人为主体的实际实验统计分析的内部和外部有效性提出了威胁。

11.2.1 对内部有效性的威胁

对一项随机化控制实验内部有效性的威胁包括随机化的失败、没有遵守处理协议、损失、实验效应和小样本容量。

随机化的失败。处理组和控制组的随机分配是随机化控制实验的基本特征,它使估计因果效应成为可能。如果处理不是随机分配的,相反,而是部分地以主体的特征或偏好为基础,那么实验的结果反映的既是处理效应,也是非随机分配效应。例如,假设在一项工作培训项目实验中,参加者被分配到处理组的依据是他们的姓是否排在前一半字母中或后一半字母中。由于姓的不同在一定程度上反映了种族的差异,因此在处理组和控制组之间种族的差别可能系统地不同。由于种族之间在工作经验、受教育程度以及其他劳动力市场特征方面是不同的,因此,在这些能够影响实验结果的遗漏因素中,处理组和控制组之间可能存在系统差异。



更一般地说,非随机化分配会导致处理 X_i 和误差项之间的相关,因为是否接受处理部分地是由进入误差项的个体特征决定的。因此一般来说,非随机化分配会使差分估计量产

在偏差,需要根据实验所评估的对象以及如何执行实验的细节进行判断。

小样本　由于以人为主体的实验可能是很昂贵的,因此有时样本容量很小。一个小容量的样本不会使因果效应的估计量产生偏差,但是这确实意味着因果效应不能被精确地估计。

一般兴趣框

Hawthorne 效应

在20世纪20年代和30年代期间,通用电气公司在他们的Hawthorne工厂进行了一系列关于工人生产效率的研究。在一组实验中,他们通过改变电灯瓦特数来了解灯光如何影响女工装备电器部件的生产效率。在其他一些实验中,他们增加或减少了休息时间,改变了车间设计,缩短了工作日。关于这些研究的一些早期很有影响的结论得出,不论灯光是更强还是更弱,劳动日是更长还是更短,工作条件是改善还是恶化,生产效率都继续上升。研究人员断言,生产效率的提高不是因为车间的变化,而是因为在实验中工人的特殊角色使他们感觉受到重视和有价值,所以他们工作越来越努力,这才导致了生产效率的提高。过了许多年,自身在一项实验中会影响主体的行为这一思想,逐渐变成了Hawthorne效应。

但是,这个故事还有个问题:仔细研究实际的Hawthorne数据表明, Hawthorne效应并不存在(Gillespie,1991;Jones,1992)! 在一些实验中,尤其在主体对结果有重要影响的实验中,仅仅处于一项实验之中就会影响其行为。更一般地说,Hawthorne效应和实验效应对内部有效性造成了威胁,即使在最初的Hawthorne数据中,Hawthorne效应并不明显。

11.2.2　对外部有效性的威胁

对外部有效性的威胁,减弱了将研究的结论推广到其他总体和环境中去的能力。有两个这样的威胁:一个是当实验样本不是所研究总体的代表时,会产生这样大的威胁;另一个是被研究的处理不是将被更广泛执行的处理的代表时,也会产生这样的威胁。

非代表性样本。所研究的总体和感兴趣的总体必须足够地相似,以保证实验结论的推广是合理的。如果一个工作培训项目实验中含有刑满释放人员,那么对这样的项目进行评估,将所研究的结论推广到其他的刑满释放人员中是可能的。然而,由于潜在的雇主对犯罪记录看得很重,因此该项结论也许不宜推广到从未犯罪的工人中。

非代表性样本的另一个例子是,当实验参加者是志愿者时,就会出现非代表性样本的情形。即使志愿者是被随机地分配到处理组和控制组中,这些志愿者也可能比整个总体的成员更积极,而对他们而言,处理可能会产生更大的效应。更一般地说,从一个更大的我们所感兴趣的总体中非随机地选择样本,可以减弱我们将从所研究的总体(例如志愿者)中得出的结论推广到我们感兴趣的总体中的能力。

非代表性项目或政策。我们所感兴趣的政策或项目也必须与我们所研究的项目足够地相似,以允许我们将结论正确地推广。一个重要的特征是,小规模的、被严格监管的实验项目,可能会非常不同于将来实际执行的项目。如果实际执行的项目是可以广泛得到的,那么一个规模将增大的项目不可能提供与实验形式相同的质量控制,或者只能在一个较低的水平上得到资助。当一个完全规模的项目比一个较小规模的实验项目效率更低时,也会产生同样的可能性。一项实验项目和一项实际项目之间的另一个差异是它们的持续时间:实验项目只持续实验期那么长,而所考虑的实际项目可能会持续更长的时间。

一般均衡效应。与规模和持续时间有关的一个问题,涉及经济学家们所称的"一般均



衡"效应。把一个小的、临时性的实验项目变成一个广泛的、持久的项目,可能要完全地改变实验结论不能被推广的那些经济环境。例如,一个规模小的、实验性的工作培训项目可能由雇主来提供,但是如果该项目变得广泛且可获得,那么它会取代雇主所提供的培训,从而降低了该项目的净利益。同样,一项普遍实行的教育改革,例如学校大幅度降低班级规模,能够增加教师的需求并改变被吸引去教书的人的类型,因此该项全面改革最终的净效应将反映在学校人员的这些被迫的变化上。用经济计量学的术语表达,在保持市场或政策环境不变的情况下,一项内部有效的小实验可能会正确地测量因果效应,但是一般均衡效应意味着,当该项目被广泛地执行时,这些其他的因素实际上并不是保持不变的。

处理和资格效应。由于在更一般的经济学和社会项目中,一个实际(非实验的)项目的参加者通常是自愿的,因此出现了对外部有效性的另一个威胁。这样,当实际执行项目的参加者被允许决定自由参加时,测度该项目对随机选择的总体成员效应的一项实验研究,一般不会提供对该项目效应的一个无偏估计量。一个工作培训项目可能对已选择参加的少数人是很有效的,而对从总体中随机选择的成员来说可能是相对无效的。解决这个问题的一个方法是,把所要做的实验设计成尽可能地与该项目将被执行的现实相吻合。例如,如果现实中的工作培训项目对于符合某一收入标准值的任何个人来说均可获得,那么该项实验协议可采用一个相似的规则:被随机选择的处理组会被给予符合项目资格的"处理",而控制组由于不符合项目资格而不被给予"处理"。在此情形下,所得的差分估计量将会估计出符合项目资格那些人的效应,它和从这样的一个工作培训项目中所估计出的处理效应是不同的,该项目是从一个够资格的总体中随机地选择成员。

11.3 使用实验数据的因果效应的回归估计量

在带有二元处理变量的一个理想的随机化控制实验中,因果效应可以用差分估计量来估计,即由公式(11.1)中β_1的OLS估计量进行估计。如果处理是被随机地接受的,那么差分估计量就是无偏的,但它却不一定是有效的。此外,如果在11.2节中所讨论的一些实际实验中的问题出现了,那么X_i和u_i是相关的,因此$\hat{\beta}_1$是有偏的。

本节提出了其他一些用于分析实验数据的基于回归分析的方法,目的就是当处理被随机地接受时,获得比差分估计量更有效的估计量;且当内部有效性的某种威胁因素存在时,获得一个无偏的或至少是一致的因果效应估计量。本节以如何检验随机化的讨论作为结束。

11.3.1 带有额外回归因子的差分估计量

通常可获得与决定实验结果有关的主体其他特征的数据。例如,由于收入依赖于以前的教育,因此在一项实验性的工作培训项目中,收入评价既依赖于以前的教育,又依赖于工作培训项目本身。在一项医药检验中,除药物治疗本身之外,健康的结果可能会依赖于病人的特征,例如年龄、体重、性别以及先前存在的医疗条件。假设$W_{1i},\cdots,W_{ri}$代表测度样本中第i个个体的r个个别特征的变量,其中,这些个别的特征并不受处理的影响(参加工作培训项目并不改变你以前的教育程度)。除了处理X_i之外,如果这些个体特征是决定结果Y_i的一个因素,那么这些变量隐含在公式(11.1)的误差项中。因此,可以对公式(11.1)进行修改,以使这些特征明确地进入到回归中。假设这些特征以线性方式进入,这就生成了多元回归模型:



$$Y_i = \beta_0 + \beta_1 X_i + \beta_2 W_{1i} + \cdots + \beta_{1+r} W_{ri} + u_i, \quad i = 1, \cdots, n \tag{11.2}$$

公式(11.2)中 β_1 的 OLS 估计量就是**带有额外回归因子的差分估计量**(difference estimator with additional regressors)。

带有额外回归因子的差分估计量的一致性。如果多元回归的四个最小二乘假设成立(见重要概念 5.4),那么公式(11.2)中所有系数的 OLS 估计量都是无偏的且一致的,而且构成了统计推断的有效基础。

在一些应用中,第一个最小二乘假设 $E(u_i | X_i, W_{1i}, \cdots, W_{ri}) = 0$ 并不是必需的。例如,如果 W 的回归因子之一是以前的教育,那么它可能与进入 u_i 中的难以观测的个人能力相关。不过,在一个比通常的条件零均值假设更弱的假设条件下,带有额外回归因子的差分估计量是一致的。具体来说,这个更弱的假设又被称为**条件均值独立性**(conditional mean independence)假设,它的数学表达式在附录 11.3 中给出。简而言之,条件均值独立性要求,给定 X_i 和变量 W 的条件下,u_i 的条件均值不依赖于 X_i,尽管它可能依赖于变量 W。

有两种重要的情形,其中,尽管在变量 W 和 u_i 之间有相关关系,但是条件均值独立性成立,而且带有额外回归因子的差分估计量是一致的(即使 W 变量系数的 OLS 估计量通常是不一致的)。第一种情形是当处理被随机地分配时,X_i 和所有的个体特征不相关,要么包括在回归(变量 W)中,要么排除在回归之外(在误差项中),因此,X_i 不能"截取"任何个体特征的效应,不论个体特征是包括在回归中还是不包括在回归中。第二种情形是已知 W_i,X_i 被有条件地随机分配时,X_i 被随机地分配,但是否被分配到处理组中的概率依赖于 W_i。例如,假设一个工作培训项目中的参加者被分成两组,高中毕业的参加者分为一组,高中没毕业的参加者分为一组。在高中毕业的参加者中,有 30% 的人被随机地分配到处理组,但是在高中没毕业的参加者中,70% 的人被随机地分配到处理组。由于每个毕业生有相同的机会被分配到处理组,因此对处理组和控制组中的高中毕业生而言,u_i 的均值是一样的。同样道理,对处理组和控制组中的非高中毕业生而言,u_i 的均值也是一样的。不过,u_i 的均值对高中毕业生和非高中毕业生通常会不同(毕业与遗漏变量即能力和动机相关)。在这种情况下,X_i 是条件随机的(给定毕业状态 W_i 条件下 X_i 被随机地分配)。如果 X_i 是条件随机的,那么就如在附录 11.3 中所进一步讨论的,条件均值独立性成立,而且具有额外回归因子的差分估计量是一致的。



重要的是,公式(11.2)中的回归因子 W_i 不是实验性结果;否则,W_i 是内生的。例如,假设 Y_i 是在工作培训项目之后的收入,W_i 代表在该项目结束之后得到的一项工作,而 X_i 表示处理。如果未来的就业状态被包含在这个回归中,那么 X_i 的系数就不再是该项目的效应了,更确切地说,在保持未来就业不变的情况下,它是该项目的局部效应。此外,未来的就业可能与 X_i 相关(该项目导致获得了工作),也可能和误差项相关(更有能力的受培训人员得到了工作)。因此,在公式(11.2)中,我们将注意力集中在变量 W 上,该变量测量了预先处理的特征,而预先处理的特征不受实验性处理的影响。

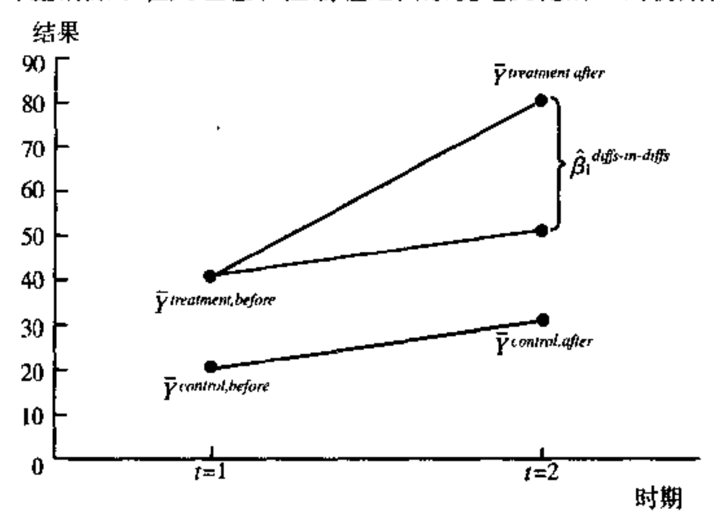
带有额外回归因子时使用差分估计量的理由。使用这个统计量存在三个理由:

1. 有效性。如果处理被随机地分配,那么,多元回归模型(公式(11.2))中 β_1 的 OLS 估计量比单个回归因子模型(公式(11.1))中的 OLS 估计量更有效(有更小的方差)。理由是,公式(11.2)中引入额外的决定 Y 的因素降低了误差项的方差(见练习 16.7)。



2. 随机化的检查。如果处理不是被随机分配的,尤其是以与 W 有关的方式分配的,那么差分估计量(公式(11.1))则是不一致的,而且一般来说与具有额外回归因子的差分估计量(公式(11.2))的概率极限也不相同。这两个 OLS 估计值之间较大的差异表明,X_i 实际

(11.4)中 $E(u_i|X_i)=0$,那么相应的差分估计量是有偏的,但差分再差分估计量却不是有偏的。图11—1就说明了这一点。在这个图中,在实验前处理组 Y 的样本均值是40,而该控制组 Y 的预先处理样本均值是20。在实验期间,控制组中 Y 的样本均值增加到30,而处理组中 Y 的样本均值增加到80。因此,处理后样本平均数的均值之差是 $80-30=50$。然而,某些差异是由于处理组和控制组有不同的预先处理均值而产生的:处理组开始就领先于控制组。例子中所得的差分再差分估计量测量了相对于控制组来说处理组的增量,在这个例子中它是 $(80-40)-(30-20)=30$。更一般地说,通过集中研究实验期间 Y 的变化,所得的差分再差分估计量消除了在处理组和控制组之间系统地变化的 Y 的初始值的影响。



注:处理组和控制组处理后之差是 $80-30=50$,但是这夸大了处理效应,因为在处理之前,处理组的 $\bar{Y}$ 比控制组高 $40-20=20$。差分再差分估计量是最终的和初始的差距之差,因此 $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}=(80-30)-(40-20)=50-20=30$。同理,差分再差分估计量是处理组的平均变化减去控制组的平均变化,即 $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}=\Delta\bar{Y}^{treatment}-\Delta\bar{Y}^{control}=(80-40)-(30-20)=30$。

图11—1 差分再差分估计量

带有额外回归因子的差分再差分估计量。差分再差分估计量可被扩展到包括额外的回归因子 $W_{1i},\cdots,W_{ri}$,这些回归因子测量了实验前的个体特征。例如,在一个工作培训项目的评价中,Y 是收入,变量 W 之一可以是参加者前置的教育水平。这些额外的回归因子可用多元回归模型来把它们放到一起刻画:

$$\Delta Y_i=\beta_0+\beta_1X_i+\beta_2W_{1i}+\cdots+\beta_{1+r}W_{ri}+u_i,\qquad i=1,\cdots,n \tag{11.5}$$

公式(11.5)中 β_1 的OLS估计量就是我们所说的**带有额外回归因子的差分再差分估计量(differences-in-differences estimator with additional regressors)**。如果 X_i 是被随机分配的,那么公式(11.5)中 $\hat{\beta}_1$ 的OLS估计量就是无偏的。

公式(11.5)中包括额外的回归因子 W 的理由,与公式(11.2)中包括它们的三个理由是一样的。公式(11.2)只使用了处理后的数据,如果 X_i 被随机地分配,那么包括额外回归因子能够改善有效性;通过增加回归因子,还可以检查随机性;通过增加回归因子允许对条件随机性进行调整,条件随机性即依赖于可观测变量 W 的随机性。重要的是,就如在公式(11.2)的上下文中所讨论的,变量 W 不包括那些实验结果就是它们本身的变量。

公式(11.5)中变量 W 的解释不同于在具有额外回归因子的差分估计量公式(公式(11.2))中的解释。在公式(11.2)中,因为只比较处理后的结果,所以变量 W 解释了 Y_i 的



水平差异。相反，在公式(11.5)中，变量 W 解释了整个实验期间 Y_i 的变化的差异。在工作培训项目的那个例子中，公式(11.5)中的因变量是实验期间收入的变化，X_i 表示参加者是否在处理组中，而 W_i 可能是前置的教育水平。这个回归把前置教育水平因素包括进来，考虑了实验期间具有更高前置教育水平的个体倾向于有更大的收入变化的可能性，不论他们是在处理组中还是在控制组中。

差分再差分估计量推广到多期：在一些实验中，个体被观测多个时期，而不止两期。在一项工作培训项目实验中，个体的收入和就业状况可能按月度被观测了一年或多年。在此情形下，以单个预先处理观测值和单个处理后观测值之间结果的变化为基础的公式(11.4)和公式(11.5)中的总体回归模型是不适合的。然而，这样的数据可用 8.3 节的固定效应回归模型进行分析，附录 11.2 提供了分析的细节。

11.3.3 不同组的因果效应估计

因果效应在不同主体之间可能是不同的，这依赖于个体的特征。例如，降低胆固醇的药对胆固醇水平的影响对高胆固醇含量的病人可能要比对低胆固醇含量的病人大。同样道理，一个工作培训项目对女人可能比对男人更有效，而且它对更积极的主体可能比对不太积极的主体更有效。更一般地说，因果效应会依赖于一个或多个变量的值，这些变量可能被观测到(如性别)，也可能不被观测到(如积极性)。

依赖于一个可观测变量值(如 W_i)的因果效应，可通过分析处理变量 X_i 与公式(11.2)(具有额外回归因子的差分估计量)中或公式(11.5)(具有额外回归因子的差分再差分估计量)中 W_i 之间的交互作用进行估计。例如，如果 W_i 是二元变量，那么这个相互作用的设定允许我们对与两个不同的 W_i 值相对应的两个不同组的处理效应进行估计。更一般地说，依赖于一个可观测回归因子的因果效应的估计，是 6.3 节中我们所讨论的交互作用方法的一个应用。

当因果效应依赖于一个不可观测的变量值时，解释因果效应估计值的讨论主题将在 11.7 节中进行研究。

11.3.4 存在局部依从性时的估计

如果存在对所用实验协议的局部依从性，那么处理水平 X_i 可能与那个不可观测的个体特征 u_i 相关，而且到目前为止我们所讨论的 OLS 估计量都是不一致的。例如，如果只有最积极的那个接受培训的人出席了工作培训项目，那么该工作培训项目可能看上去是有效的，但只因为受培训的人是最努力的工人，所以不管他们是否参加该培训项目，他们在劳动力市场上都会做得很好。

如在第 10 章中所讨论的，假设存在一个可获得的工具变量，那么工具变量回归为一个回归因子与误差项之间的相关问题提供了一个通解。在一个具有局部依从性的实验中，分配的处理水平可作为实际处理水平的一个工具变量。

回想一下，一个变量要成为一个有效的工具变量，必须满足工具变量相关性和外生性这两个条件(见重要概念 10.3)。只要协议被部分地遵守，那么实际的处理水平(X_i)就部分地由分配的处理水平(Z_i)所决定，因此工具变量 Z_i 是相关的。如果分配的处理水平是随机决定的——也就是说，如果实验是随机分配的——而且分配除了通过对处理是否被接受产生影响外，如果该分配本身对结果没有影响，那么 Z_i 是外生的。也就是说，Z_i 的随机分配意味着 $E(u_i|Z_i)=0$，其中，u_i 是公式(11.1)差分设定中的误差项还是公式(11.4)差分再差分



设定中的误差项，依赖于使用哪个估计量。这样，在一个存在局部依从性和随机分配处理的实验中，最初的随机分配是个有效的工具变量。

11.3.5 随机化的检验

随机化可以通过检查该随机化变量实际上是否依赖于任何可观测的个体特征来检验。

检验处理的随机接受　如果处理是被随机地接受的，那么 X_i 将与这个可观测的个体特征不相关。因此，“处理是被随机地接受的”这一假设，可通过检验“在一个 X_i 对 $W_{1i},\cdots,W_{ri}$ 的回归中，$W_{1i},\cdots,W_{ri}$ 的系数都是0”这一假设进行检验。在工作培训项目那个例子中，用接受工作培训(X_i)对性别、种族和前置教育水平(W)进行回归，并计算检验 W 系数是否都是0的 F 统计量，可对以下假设提供检验，即“处理是被随机地接受的”零假设和“接受处理依赖于性别、种族或前置教育水平”的备择假设。[①]

检验随机分配。如果处理被随机地分配，那么该分配 Z_i 将与可观测的个体特征不相关。因此，“处理被随机地分配”这个假设可通过将 Z_i 对 $W_{1i},\cdots,W_{ri}$进行回归以及检验所有的斜率系数都是0的零假设进行检验。

11.4 减小班级规模效应的实验估计值

在这一节，我们回到第2部分中所处理的问题：减小低年级班级的规模对考试成绩的效应是什么？在20世纪80年代后期，田纳西州执行了一个很大的、花费几百万美元的随机化控制实验，以确定减小班级规模是否是一种改善初等教育的有效方法。这个实验的结论已强烈地影响到我们对减小班级规模效应的理解。

11.4.1 实验设计

田纳西州减小班级规模的实验，即所知的 STAR 项目(Student-Teacher Achievement Ratio)，它是用来评估小班对学习影响的一个四年期的实验。由田纳西州立法委员会提供基金，在这四年间这个实验大约花费了1 200万美元。该项研究比较了从幼儿园直到三年级三种不同的班级安排：常规班，每班22～25个学生，一位教师，没有助教；小型班，每班13～17个学生，没有助教；还有一种班型，即一个常规班，再加上一位老师的助教。

参与该实验的学校每种类型的班级至少有一个，在1985—1986学年初，进入到该学校幼儿园的学生，被随机地分配到这三种类型的班级中，教师也被随机地分配到这三种类型的班级中。

根据最初的实验协议，学生在他们最初所分配的班级中需要呆四年，也即实验期的四年(从幼儿园直到三年级)，然而，由于学生家长的抱怨，在一年级刚开始，最初被分配到常规班(有助教或没有助教)的学生被随机地重新分配到有助教的常规班或没有助教的常规班中；最初被分配到小班的学生仍留在小班里。进入学校的一年级新生(幼儿园是可以任意选择的)，在实验的第二年被随机地分配到这三种类型的班级中。每年，实验中的学生都要进行阅读和数学的标准化考试(斯坦福达标考试)。

为了达到目标的班级规模，这个项目对额外教师和助教给予了补偿。在该项研究的第



---

① 在这个例子中，X_i 是二元变量，因此如第9章中所讨论的，X_i 对 $W_{1i},\cdots,W_{ri}$的回归是一个线性概率模型，异方差稳健的标准误是非常必要的。当 X_i 是二元变量时，检验假设 $E(X_i|W_{1i},\cdots,W_{ri})$不依赖于 $W_{1i},\cdots,W_{ri}$的另一种方法是使用 probit 模型或 logit 模型(见9.2节)。

一年期间,大约有6 400名学生参加了108个小班、101个常规班和99个有助教的常规班。在整个研究的4年间,有80所学校共约11 600名学生参加了这项研究。

与实验设计的偏离 该实验协议规定,除了在一开始的一年级通过随机化分配以外,学生不应该换班,然而,约有10%的学生在随后的年份里因性格不合和行为举止等问题而换班。这些换班行为代表了与随机化计划的偏离,这依赖于换班的真实性质。这些换班行为存在将偏差引入结论的潜在可能性。对于那些纯粹为了避免个性冲突而进行的换班,可能与实验结果完全无关,所以他们不会引入偏差。然而,如果因为家长关心他们孩子的教育,向学校施加压力将孩子转到小班而出现换班,那么没有遵守该实验协议会使结论偏向于夸大小班的效果。与实验协议的另一种偏离是,由于学生换班和迁入迁出该学区,班级规模随时间发生了变化。

11.4.2 STAR数据分析

由于存在两个处理组——小班和有助教的常规班,因此差分估计量的回归形式需要修改,以处理两个处理组和控制组。这可通过引入两个二元变量来实现,一个变量表示学生是否在小班中,另一个变量表示学生是否在有助教的常规班中,于是生成了如下的总体回归模型:

$$Y_i = \beta_0 + \beta_1 SmallClass_i + \beta_2 RegAide_i + u_i \tag{11.6}$$

其中,如果第i个学生在小班中,那么$SmallClass_i=1$;否则,$SmallClass_i=0$。如果第i个学生在有助教的常规班中,那么$RegAide_i=1$;否则,$RegAide_i=0$。Y_i是考试成绩。相对于常规班来说,小班对考试成绩的效应是β_1;相对于小班来说,有助教的常规班的效应是β_2。该实验的差分估计量可通过用OLS估计方程(11.6)中的β_1和β_2进行计算。

表11—1给出了在小班或有助教的常规班中对考试成绩效应的差分估计值。表11—1回归中的因变量Y_i是学生合并的斯坦福达标考试的数学和阅读部分的总成绩。根据表11—1中的估计值,对幼儿园学生而言,相对于在常规班中,在小班中对考试成绩的效应是增加13.9分;在有助教的常规班中对考试成绩的估计效应是增加0.31分。对每个年级而言,小班没有提高成绩的零假设在1%(双边)的显著性水平下被拒绝。不过,相对于没有助教的常规班,拒绝有助教的常规班没有提高成绩的零假设是不可能的,除了在一年级。虽然一年级的估计值比较大,但是在幼儿园、二年级和三年级中,小班所估计的提高的幅度普遍地相似。

表11—1 STAR项目:班级规模处理组对标准化考试成绩效应的差分估计值

回归因子	年级			
	K	1	2	3
小班	13.90**	29.78**	19.39**	15.59**
	(2.45)	(2.83)	(2.71)	(2.40)
有助教的常规班	0.31	11.96**	3.48	-0.29
	(2.27)	(2.65)	(2.54)	(2.27)
截距	918 04**	1 039.39**	1 157.81**	1 228.51**
	(1 63)	(1.78)	(1.82)	(1.68)
观测期数	5 786	6 379	6 049	5 967



注:上述回归是使用附录11.1中所描述的STAR项目公共数据集估计的。因变量是在斯坦福达标考试中数学和阅读部分学生的合并成绩。标准误在系数下面的括号中给出。**代表使用双边检验时单个系数在1%的显著性水平下在统计上是显著的。

的2.16。

由于教师被随机地分到校内所有类型的班级中去，因此实验也提供了估计教师的经验对考试成绩的效应的机会。然而，教师在工作的学校之间并不是被随机地分配的，一些学校比另一些学校有更多的经验丰富的教师。因而，教师经验可能与误差项相关，如果有丰富经验的教师在具有更多资源和更高平均考试成绩的学校工作，那么情况就会如此。因此，要估计教师经验对考试成绩的效应，我们必须控制学校的其他特征，这可通过对每个学校使用一组完备的指示变量（"学校效应"）来实现，也即代表学生所上学校的指示变量。由于教师在校内是被随机地分配的，因此对于给定的学校，u_i 的条件均值并不依赖于该处理。用附录11.3中的术语表述，由于校内的随机分配，条件均值独立性假设成立，其中额外回归因子 W 是学校的效应。当学校效应被包含进来时，教师经验效应的估计值减小了一半，从第(2)列的1.74减小到第(3)列的0.74。即使这样，第(3)列的估计值在统计上仍是显著的，而且数值又适度地大。10年的经验对应于考试成绩预测值增加7.4分。

现在我们来尝试解释表11—2中的一些其他系数。例如，在这些标准化考试中，幼儿园里的男孩比女孩成绩差，但是，这些个体学生的特性并不是被随机分配的（参加考试的学生性别并不是被随机分配的!），所以这些额外回归因子可能与遗漏变量相关。例如，如果种族因素或免费享受午餐资格与减少的校外学习的机会相关（表11—2的回归中遗漏了这个），那么它们的估计系数会反映这些遗漏的影响。如11—2节中所讨论的，如果处理被随机地分配，那么其系数的估计量都是一致的，不论其他回归因于是否与误差项相关。但是，如果额外回归因子与误差项相关，那么它们的系数估计量具有遗漏变量偏差。

解释班级规模的估计效应。从实际意义上说，表11—1和表11—2所给出的班级规模的估计效应是大还是小呢？回答这个问题有两种方法：第一，将原始考试成绩的估计变化转化为考试成绩的标准差单位，这样，表11—1中的估计值在年级之间都是可比的；第二，将所估计的班级规模的效应和表11—2中的其他系数进行比较。

由于每个年级考试成绩的分布都不是一样的，因此表11—1中所估计的效应在年级间并不是直接可比的。我们在7.3节中就面对过这个问题，当时，我们想要比较使用取自加利福尼亚州的数据所估计的减小学生—教师比对考试成绩的效应和以马萨诸塞州的数据为基础的估计值。由于这两个检验是不同的，因此系数不能直接作比较。7.3节中的解决方法是，将所估计的效应转化为考试成绩的标准差单位，因此学生—教师比的单位减少，对应于所估计的考试成绩标准差的变化。在此，我们采用这种方法，目的是使表11—1中的估计效应在年级之间是可比较的。例如，幼儿园里孩子的考试成绩标准差是73.7分，因此，依据表11—1中的估计值，在幼儿园小班中的效应是13.9/73.7 = 0.19，标准误是2.45/73.7 = 0.03。将表11—1中班级规模的估计效应，转化成学生之间考试成绩的标准差单位，其结果在表11—3中做了总结。如用标准差单位来表达，在小班中的估计效应对幼儿园、二年级以及三年级而言是相似的，大约为考试成绩一个标准差的1/5。同样，对幼儿园、二年级和三年级而言，在一个带有助教的常规班中的估计效应约为0。对一年级来说，估计的处理效应比较大。然而，对一年级而言，小班和带有一个助教的常规班之间的估计差异是0.20分，与其他年级一样。因而，对一年级的结果的一个解释是，控制组中的学生——不带有助教的常规班——在那一年由于一些不同寻常的原因，也许仅仅是由于简单随机抽样的变差，而碰巧在考试上表现得较差。



另一种测量小班估计效应大小的方法是，比较所估计的处理效应和表11—2中的其他系数。在幼儿园里，被分配在小班中对考试成绩的估计效应是13.9分（表11—2的第(1)

行)。保持种族、教师的工作年限、免费享受午餐的资格以及处理组不变,根据表 11—2 的第(4)列中的估计值,在标准化考试中,男生成绩比女生成绩约低 12 分。因此,被分配在小班中的估计效应要比女生和男生之间的成绩差距稍大一些。作为另一种比较,第(4)列中教师工作年限的估计系数是 0.66,所以,一位具有 20 年教龄的教师估计会使考试成绩提高 13 分。因而,相对于新教师而言,在小班中的估计效应与具有 20 年教龄的富有经验的教师的效应大概是一样的。这些比较表明了在小班中的估计效应是巨大的。

表 11—3　　**用考试成绩的标准差单位表示的班级规模的估计效应**

处理组	年级			
	K	1	2	3
小班	0.19**	0.33**	0.23**	0.21**
	(0.03)	(0.03)	(0.03)	(0.03)
有助教的常规班	0.31	0.13**	0.04	0.00
	(2.27)	(0.03)	(0.03)	(0.03)
考试成绩标准差(s_Y)	73.70	91.30	84.10	73.30

注:前两行中的估计值和标准误是表 11—1 中的估计效应,除以该年级的斯坦福考试成绩的样本标准差(表 11—1 中的最后一行),使用实验中的学生数据所计算得到的。标准误在系数下面的括号中给出。**代表使用双边检验时个别系数在 1% 的显著性水平下在统计上是显著的。

其他一些结论。经济计量学家、统计学家以及初等教育专家已研究了这个实验的不同方面,在此我们简要地总结其中的部分研究发现。发现之一是,小班的效应集中在最低的年级,这从表 11—3 中可以看出,除了一年级有异常结论之外,表 11—3 中给出的常规班和小班之间考试成绩的差距基本上在年级之间是个常数(幼儿园是 0.19 个标准差单位,二年级是 0.23,三年级是 0.21)。由于最初被分到小班的孩子仍留在那个小班中,因此,这意味着留在小班中不会产生额外好处,但是,最初的分配所获得的好处由高年级所获得,不过在处理组和控制组之间的差距没有增加。另一个研究发现是,如表 11—3 的第(2)行所示,这个实验表明在常规班中配备助教几乎没有什么好处。关于解释该实验结果的一个潜在理由是,一些学生没有遵守该处理协议(一些学生从小班中转出)。如果在幼儿园中的最初安置是随机的,而且对考试成绩没有直接影响,那么最初安置可被作为一个部分地而不是完全地影响安置的工具变量。这个方法最早是由 Krueger(1999)提出来的,利用最初的班级安置作为工具变量,他使用两阶段最小二乘(TSLS)估计了班级规模对考试成绩的效应。他发现 TSLS 和 OLS 估计值是相似的,这导致他得出这样的结论:与实验协议的偏离并没有对 OLS 估计值产生巨大的偏差。①

11.4.3　班级规模效应的观测估计值和实验估计值比较

基于加利福尼亚州和马萨诸塞州学区的观测数据,第 2 部分给出了班级规模效应的多元回归估计值。在那些数据中,班级规模并不是被随机分配的,而是由地方学校的官员在平衡理想的教育目标和现实的预算约束条件下决定的。怎样把这些观测估计值与 STAR 项目



① 为进一步阅读 STAR 项目的资料,请见 Mosteller(1995),Mosteller, Light 与 Sachs(1996)和 Krueger(1999)。Ehrenberg,Brewer,Gamoran 与 Willms(2001a,2001b)讨论了 STAR 项目,并在政策争论的意义下,把讨论放在班级规模和这个主题的有关研究上。对 STAR 项目的一些批判,请见 Hanushek(1999a)。对班级大小和考试成绩之间关系更一般的重要观点,请见 Hanushek(1999b)

的实验估计值相比较呢?

为了将加利福尼亚州和马萨诸塞州的估计值同表11—3中的估计值比较,这样做是非常必要的,即以相同的班级减小规模并用考试成绩的标准差单位来表示预测效应。在STAR项目实验的4年间,小班比大班平均约少7.5个学生,因此我们使用观测估计值来预测每班减少7.5个学生对考试成绩的效应。根据表7—3第(1)列中所总结的线性设定的OLS估计值,加利福尼亚州的估计值使我们预测出,学生—教师比减少7.5个单位会使考试成绩增加5.5分(0.73×7.5)。加利福尼亚州学生间的考试成绩标准差大约是38分,因此,用学生间考试成绩的标准差单位来表示,减少7.5个学生的估计效应是5.5/38≈0.14个标准差。[①] 加利福尼亚州所估计的斜率系数的标准误是0.26(见表7—3),因此以标准差单位表示的减少7.5个学生的估计效应的标准误是0.26×7.5/38≈0.05。这样,基于加利福尼亚州的数据,用学生间考试成绩标准差单位表示,每班减少7.5个学生的估计效应是0.14个标准差,标准误是0.05。这些计算以及马萨诸塞州的类似计算,连同取自表11—2第(1)列中对幼儿园的STAR估计值一起,在表11—4中总结。

表11—4　　基于STAR数据和加利弗尼亚州与马萨诸塞州的观测数据计算的,学生—教师比减少7.5的估计效应

研究项目	$\hat{\beta}_1$	学生—教师比变化	学生考试成绩的标准差	估计效应	95%的置信区间
STAR (年级K)	-13.90^{**} (2.45)	小班与常规班	73.8	0.19^{**} (0.03)	(0.13,0.25)
加利福尼亚州	-0.73^{**} (0.26)	-7.5	38.0	0.14^{**} (0.05)	(0.04,0.24)
马萨诸塞州	-0.64^{*} (0.27)	-7.5	39.0	0.12^{*} (0.05)	(0.02,0.22)

注:STAR研究的估计系数$\hat{\beta}_1$取自表11—2的第(1)列。加利福尼亚州和马萨诸塞州研究的估计系数$\hat{\beta}_1$取自表7—3的第(1)列。所估计的效应是指相对于常规班来说在小班中的效应(对于STAR),或者说是学生—教师比减少7.5个单位的效应(对加利福尼亚州和马萨诸塞州来说的)。减少学生—教师比的95%的置信区间是这个估计效应±1.96倍标准误。标准误在估计效应下面的括号中给出。当使用双边检验时,所估计的效应在*5%的显著性水平下或在**1%的显著性水平下在统计上是显著地异于0的。

从加利福尼亚州和马萨诸塞州的观测研究中得出的估计效应比STAR的估计效应稍微小一些。然而,解释不同研究的估计值不相同的一个原因是随机抽样的变化性,因此,比较三个研究所得出的估计效应的置信区间是有意义的。基于幼儿园的STAR数据,在小班中效应的95%的置信区间(在表11—4的最后一列中给出)是0.13到0.25。基于加利福尼亚州的观测数据,类似的95%的置信区间是0.04到0.24,而对马萨诸塞州来说,相应的95%的置信区间是0.02到0.22。因而,从加利福尼亚州和马萨诸塞州的研究中得出的95%的置信区间包含了从STAR幼儿园数据中得出的95%的置信区间的大部分。从这个方面来看,这三个研究都给出了普遍相似的估计值范围。

有许多理由可以解释为什么实验估计值与观测估计值不同。一个理由是,如7.3节中



① 在表7—3中,估计效应是按照地区间的考试成绩标准差给出的,在表11—3中,估计效应是根据学生间的考试成绩标准差给出的。学生间的标准差大于地区间的标准差,对加利福尼亚州而言,学生间的标准差是38,但地区间的标准差是19.1。

所讨论的，仍然存在对该观测性研究内部有效性的残余威胁。例如，由于经常有孩子迁出迁进该地区，该地区的学生—教师比可能并不反映学生实际所经历的学生—教师比，因此，由于存在变量误差偏差，加利福尼亚州和马萨诸塞州研究中的学生—教师比的系数可能会偏向于0。其他的理由主要关注外部有效性。在观测研究中所使用的地区平均的学生—教师比与班级中的实际孩子数(STAR项目中的实验变量)不是同一回事。STAR项目发生在20世纪80年代的南部州，可能不同于1998年的加利福尼亚州和马萨诸塞州，且所比较的年级不同(在STAR中是K~3年级，在马萨诸塞州是4年级，在加利福尼亚州是5年级)。就所有这三个预测不同估计值的理由来说，这三个研究的结论是非常相似的。观测性研究类似于STAR项目的估计值这一事实表明，观测估计值内部有效性的残余威胁是很小的。

11.5 准实验

真正的随机化控制实验是很昂贵的——STAR实验花费了1 200万美元——而且它们常会引发道德问题。在医学中，通过随机地将主体分配到吸烟的处理组中和不吸烟的控制组中，以确定吸烟对寿命的影响，这是不道德的。在经济学中，通过向随机选择的高中生出售带有补贴的香烟来估计青少年对香烟的需求弹性也是不道德的。出于成本、道德和现实的原因，经济学中真正的随机化控制实验很少。

不过，随机化控制实验的统计观点和方法可以被推广到非实验的环境中来。在一个**准实验**(quasi-experiment)，也即**自然实验**(natural experiment)中，随机性是被在个体的环境中发生的变差引入的，因此处理仿佛是被随机地分配的。这些在个体环境中发生的变差可能由多种原因所引发，例如，法律机构反复无常的行为，场所的变化，政策或项目执行的时效性，自然随机性；例如，出生日期，降雨量，或与当前所研究的因果效应无关的其他因素等。

存在两种类型的准实验。在第一种类型的准实验中，不论一个个体(或更一般地说，一个实体)是否接受处理，都被认为仿佛是被随机地决定的。在此情形下，利用该处理X_i作为回归因子，可以用OLS估计所研究的因果效应。

在第二种类型的准实验中，所谓“仿佛”随机变化只是处理的部分决定性因素。11.3节讨论了在一项实验中，当随机分配影响到实际接受的处理时，该随机分配可被作为一个工具变量来使用。同理，在一项准实验中，“仿佛”随机变化有时提供了一个可影响实际接受的处理(X_i)的工具变量(Z_i)。因此，用工具变量回归估计因果效应，这里变差的“仿佛”随机性来源提供了该工具变量。

11.5.1 几个例子

我们用例子来说明这两种类型的准实验。第一个例子是一个准实验的例子，其中的处理是“仿佛”被随机地决定的。在第二个和第三个准实验的例子中，“仿佛”的随机变差影响了(但并不完全决定)处理的水平。

例1：劳动力市场中的移民效应。移民会降低工资吗？经济学理论认为，如果劳动力供给由于移民流入而增加，那么劳动力的“价格”即工资将会下降。然而，在所有其他条件都相等的情况下，移民被吸引到劳动力需求高的城市，因此，移民对工资效应的OLS估计量将会是有偏的。一个理想的估计移民对工资效应的随机化控制实验，将把不同数量的移民(不同的“处理”)随机地分配到不同的劳动力市场(“主体”)，并测度对工资的效应(“结果”)。不过，这样的实验面临着严重的实际、财务和道德的问题。



因此，劳动经济学家 David Card(1990)使用了一个准实验，在这个准实验中，在 1980 年临时取消对古巴移民限制的“玛丽亚尔港船只解封(Mariel boatlift)”期间，大量古巴移民进入佛罗里达州的迈阿密劳动力市场。半数移民定居在迈阿密，部分是由于迈阿密有个大的先前存在的古巴社区。通过比较迈阿密的不熟练工人的工资变化和同期美国其他相似城市中同类工人的工资变化，Card 使用差分再差分估计量来估计移民增加的因果效应。他得出结论，移民的流入对不熟练工人工资的效应是可以忽略的。

例 2：服兵役对平民收入的效应。在军队中服兵役会改善你在劳动力市场上的前景吗？军队提供培训，这种培训对未来的雇主来说可能是有吸引力的。然而，由于服兵役至少部分地是由个人选择和特征决定的，因此，通过个体平民的收入对以前是否服过兵役的 OLS 回归可能会得到一个服兵役对平民收入效应的有偏估计量。例如，军队只接受那些满足最低体检要求的申请人，而在私人部门的劳动力市场中，混得不好可能更会使一个人报名参军。

为了防止这个选择偏差的发生，Joshua Angrist(1990)使用了一个准实验设计，在这个准实验设计中，他研究了越战期间在美国军队中服过役的人的劳动力市场的历史纪录。在这个期间，一个年轻人是否被应征入伍，部分地是由人的生日的全国性抽彩系统决定的：年轻人被随机地分配，取得低的抽彩号的人有资格应征入伍，而取得高的抽彩号的人不能入伍，但实际能否入伍则是由复杂的规则决定的，包括体检筛选和某种豁免，而且一些年轻人志愿参军，因此，在军队中服役只是部分地受到你是否符合应征条件的影响。这样，将符合应征条件作为一个工具变量，该工具变量部分地决定是否服兵役，但又是被随机分配的。在这种情况下，存在通过抽彩来分配符合应征资格这样一个真正的随机分配，但是因为这种随机化没有在评估服兵役效应的实验中被充分地采用，所以这是一个准实验。Angrist 的结论是，服兵役的长期效应是减少白人老兵的收入，但不会减少非白人老兵的收入。

例 3：心肌导管插入术的效应。10.5 节描述了由 McClellan，McNeil 和 Newhouse(1994)所做的研究。在该项研究中，他们使用从心脏病人的家到心肌导管插入术医院的距离(相对于从心脏病人的家到没有心肌导管插入术设备的医院的距离)，作为心肌导管插入术实际处理的工具变量。这项研究就是一个准实验，其中含有一个部分地决定处理的变量。处理本身，也即心肌导管插入术，是由病人的个人特征以及病人和医生的决策决定的，然而，它也受附近的医院是否有能力进行这个手术的影响。如果这个病人的位置“仿佛”是被随机地分配的，且对健康结果没有直接影响，而不是通过它的影响来改变心肌导管插入术的概率，那么到一所心肌导管插入术医院的相对距离就是一个有效的工具变量。

其他的例子。准实验研究策略也已被应用到其他研究领域。Garvey 与 Hanka(1999)利用美国州法律的变差来研究反接管法对公司财务结构(例如，公司债务的使用)的效应。Meyer，Viscusi 与 Durbin(1995)使用肯塔基州和密歇根州丰饶的失业保险收益(它有差别地影响高收入工人而不是低收入工人)的大的离散变化来估计失业收益变化对失业时间的影响。Meyer(1995)，Rosenzweig 与 Wolpin(2000)以及 Angrist 与 Krueger(2001)的研究给出了经济学和社会政策领域中准实验应用的其他例子。

11.5.2 分析准实验的经济计量方法



分析准实验的经济计量方法大致上与 11.3 节中所列出的分析真正实验的方法是一样的。如果处理水平 X 是“仿佛”被随机决定的，那么 X 系数的 OLS 估计量是因果效应的一个无偏估计量。如果处理水平仅仅部分地是随机的但受到一个“仿佛”是被随机分配的变量 Z 的影响，那么该因果效应的估计可通过将 Z 作为一个工具变量的工具变量回归来完成。

由于准实验一般地没有真正的随机化，因此在处理组和控制组之间可能会存在系统差异。如果是这样的话，那么在这个回归中引入单个主体预先处理特征的可观测的测量指标是非常重要的（即 11.3 节回归中的 W）。如同 11.3 节中所讨论的，一般地，引入内生的（即处理的结果）回归因子 W 会引致因果效应估计量的不一致。

在准实验中，数据的收集一般并不是仅仅为了某一特定研究的需要，所以这样的准实验“主体”的面板数据有时是得不到的（一个例外情况在方框中的关于最低工资的例子中做了介绍）。如果是这样的话，进一步研究的方法就是使用随时间所搜集的一系列截面数据，并根据对重复截面数据的使用修改 11.3 节中的方法。

使用重复截面数据的差分再差分。**重复截面数据**（repeated cross-sectional data）集是截面数据集的集合，其中每个截面数据集对应着一个不同的时期。例如，某数据集可能包含 400 个不同个体在 2001 年的观测值，500 个不同个体在 2002 年的观测值，总共 900 个不同的个体。政治投票数据是重复截面数据的一个例子，在这个数据中，政治偏好用一系列随机选择的潜在投票者的调查进行测量，这里的调查是在不同的时期进行的，而且每次调查都有不同的回答者。

使用重复截面数据的前提条件是，如果个体（更一般地讲是实体）是从某一相同的总体中随机地抽取的，那么在早期截面中的个体可被用来作为后期截面中处理组和控制组中个体的代替物。例如，假设与地方劳动力市场没有关系的一项基金增加了，那么一个工作培训项目计划在南加利福尼亚州被扩大了，而不是在北加利福尼亚州。假设你有两个随机选择的加利福尼亚州成年人的截面数据，其中一个调查是在培训计划扩张前实施的，另一个调查是在培训计划扩张后实施的，那么“处理组”将是南加利福尼亚人，“控制组”将是北加利福尼亚人。在处理之前，你没有实际被处理的南加利福尼亚人的数据（因为你没有面板数据），但是你有与被处理的那些人在统计上类似的南加利福尼亚人的数据，这样，你可以使用第一个时期中南加利福尼亚人的截面数据作为处理组的预先处理观测值的代替物，并使用北加利福尼亚人的截面数据作为控制组预先处理观测值的代替物。

一般兴趣框

最低工资对就业的影响是什么

最低工资的增加会使对不熟练工人的需求减少多少？经济理论指出，当价格上升时，需求便下降，但确切地说数量是多少却是一个实证性问题。由于价格和数量是由供给和需求决定的，因此在就业对工资回归的 OLS 估计量中存在联立因果偏差。假设一个随机化控制实验可以将不同的最低工资随机地分配给不同的雇主，而后比较处理组和控制组就业的变化（结果），在实际中这种假设的实验如何操作呢？

劳动经济学家 David Card 和 Alan Krueger（1994）决定进行这样的实验，但是他们决定让“自然”——或更准确地说是，地理——为他们执行随机化。在 1992 年，新泽西州的最低工资从每小时 4.25 美元上升到 5.05 美元，但是邻近的宾夕法尼亚州的最低工资却保持不变。在这个实验中，最低工资增加的“处理”——被设置在新泽西州或宾夕法尼亚州——被认为“仿佛”是随机分配的，也可以这样来理解，受制于工资增加与这个时期就业变化的其他决定性因素假定不相关。Card 与 Krueger 搜集了这两个州的快餐店在工资增加前后的就业数据。当他们计算差分再差分估计量时，他们发现了一个令人惊讶的结果：相对于宾夕法尼亚州的快餐业而言，没有证据表明新泽西州的快餐业就业率下降了。实际上他们的一些估计值显示，相对于宾夕法尼亚州来说，在快餐业的最低工资上升后，新泽西州快餐店的就



业率增加了！

这个发现与基本微观经济理论相冲突，并产生了很大的争议。后面的一些分析使用了不同来源的就业数据，结果表明，在工资增加之后，新泽西州的就业可能已有一点下降，但是即使如此，所估计的劳动力需求曲线也是非常缺乏弹性的（Neumark 与 Wascher，2000）。虽然在这个准实验中，精确的工资弹性是个有争议的问题，但是最低工资的增加对就业的影响看起来比许多经济学家以前所认为的要小。

当有两个时期时，重复截面数据的回归模型是：

$$Y_{it} = \beta_0 + \beta_1 X_{it} + \beta_2 G_i + \beta_3 D_t + \beta_4 W_{1it} + \cdots + \beta_{3+r} W_{rit} + u_{it} \qquad (11.7)$$

其中，X_{it}是截面数据中第 i 个个体（实体）在时期 $t(t=1,2)$ 的实际处理，D_t 是个二元指示变量，它在第一期等于 0，在第二期等于 1。G_i 也是一个二元变量，表示某个个体是否在处理组中（或者是否在替代处理组中，如果观测值是在预先处理期）。如果他或她在第二期的处理组中，则第 i 个个体接受处理。因此，在公式（11.7）中，$X_{it}=G_i \times D_t$，也就是说，X_{it}是 G_i 和 D_t 之间的交互作用。

第3部分

如果这个准实验使 X_{it}"仿佛"是被随机地接受，那么因果效应可用公式（11.7）中 β_1 的 OLS 估计量进行估计。如果有两个以上的时期，那么公式（11.7）应被修改为包含 $T-1$ 个表示不同时期的二元变量（见附录 11.2）。

如果该准实验使处理 X_{it}只是部分地被随机接受，那么通常 X_{it}将与 u_{it}相关，OLS 估计量是有偏的且是不一致的。在此情形下，该准实验中随机性的来源采取工具变量 Z_{it}的形式，Z_{it}部分地影响处理水平，而且"仿佛"是被随机分配的。和通常一样，要使 Z_{it}变成一个有效的工具变量，它必须是相关的（也就是说，它必须与实际处理 X_{it}相关）和外生的。

11.6 准实验中存在的潜在问题

和所有的实证研究一样，准实验也面临着内部有效性和外部有效性的威胁。对内部有效性的一个特别重要的潜在威胁是，"仿佛"随机化实际上能否被可靠地看做是真正的随机化。

11.6.1 对内部有效性的威胁

在 11.2 节中所列出的对真正随机化控制实验内部有效性的威胁，也适用于准实验，但需要做一些修改。

随机化失败。准实验依赖于个体环境的差异——法律变化、突发性不相关事件等等——来提供处理水平中的"仿佛"随机化。如果这个"仿佛"随机化没有生成随机的处理水平 X（或工具变量 Z），那么一般地说，OLS 估计量是有偏的（或者该工具变量估计量是不一致的）。

和在真正的实验中一样，检验随机化失败的一种方法是，检查处理组和控制组之间的系统差异，例如，通过用 X（或 Z）对个体特征（W）回归，并检验 W 的系数都是 0 的假设。如果存在不容易被该准实验的性质所解释的差异，那么这就是该准实验没有产生真正随机化的证据。即使在 X（或 Z）与 W 之间不存在相关关系，X（或 Z）也可能与误差项 u 中的一些观测不到的因素有关。由于这些因素观测不到，因此不能被检验，"仿佛"随机化这一假设的有效性必须使用特定应用场合的专门知识和判断进行评价。

没有遵守处理协议。在一个真正的实验中，当处理组中的成员没有接受处理和/或控制



差分估计量就是该平均因果效应的一个一致估计量。该 OLS 估计量是 $\hat{\beta}_1 = s_{XY}/s_x^2$（公式(4.8)）。如果观测值是独立同分布的，那么样本协方差和方差就是总体协方差和方差的一致估计量，即 $\hat{\beta}_1 \xrightarrow{p} \sigma_{XY}/\sigma_X^2$。如果 X_i 是被随机分配的，那么 X_i 就独立分布于其他个体的特征，包括可观测到的特征和不可观测到的特征，尤其是独立分布于 β_{1i}。因此，这个 OLS 估计量 $\hat{\beta}_1$ 具有极限：

$$\hat{\beta}_1 = \frac{s_{XY}}{s_X^2} \xrightarrow{p} \frac{\sigma_{XY}}{\sigma_X^2} = \frac{\mathrm{cov}(\beta_0 + \beta_{1i}X_i + u_i, X_i)}{\sigma_X^2} = \frac{\mathrm{cov}(\beta_{1i}X_i, X_i)}{\sigma_X^2} = E(\beta_{1i}) \quad (11.9)$$

这里，第三个等式使用了重要概念 2.3 中关于协方差和 $\mathrm{cov}(u_i, X_i) = 0$ 的事实，其中，$E(u_i | X_i) = 0$（公式(2.25)）隐含着 $\mathrm{cov}(u_i, X_i) = 0$。最后一个等式是从 β_{1i} 独立分布于 X_i 得出的。如果 X_i 是被随机决定的，那么 β_{1i} 就确实独立分布于 X_i（见练习 11.7）。因此，如果 X_i 是被随机决定的，那么 $\hat{\beta}_1$ 就是平均因果效应 $E(\beta_{1i})$ 的一个一致性估计量。

11.7.3 具有异质因果效应的 IV 回归



假设处理只是部分地被随机决定的，Z_i 是个有效的工具变量（相关的和外生的），在 Z_i 对 X_i 的效应中存在异质性。具体来说，假设 X_i 与 Z_i 是以下列线性模型联系的：

$$X_i = \pi_0 + \pi_{1i}Z_i + v_i \text{[1]} \quad (11.10)$$

其中，系数 π_{1i} 随个体的不同而变化。公式(11.10)是 TSLS 的第一阶段方程（公式(10.2)），只是做了一定的修改，这个修改允许 Z_i 的变化对 X_i 的效应在不同的个体之间是变化的。

TSLS 估计量是 $\hat{\beta}_1^{TSLS} = s_{ZY}/s_{ZX}$（公式(10.4)），它是 Z 和 Y 之间的样本协方差与 Z 和 X 之间的样本协方差之比。如果这些样本观测值是独立同分布的，那么这些样本协方差就是总体协方差的一致估计量，因此，有 $\hat{\beta}_1^{TSLS} \xrightarrow{p} \sigma_{ZY}/\sigma_{ZX}$。假设 π_{1i} 和 β_{1i} 独立分布于 u_i，v_i 和 Z_i，而且 $E(u_i|Z_i) = E(v_i|Z_i) = 0$，$E(\pi_{1i}) \neq 0$。附录 11.4 中证明了。在这些假设之下有：

$$\hat{\beta}_1^{TSLS} = \frac{s_{ZY}}{s_{ZX}} \xrightarrow{p} \frac{\sigma_{ZY}}{\sigma_{ZX}} = \frac{E(\beta_{1i}\pi_{1i})}{E(\pi_{1i})} \quad (11.11)$$

也就是说，TSLS 估计量依概率收敛于 β_{1i} 和 π_{1i} 之积的期望值同 π_{1i} 的期望值之比。

公式(11.11)中最后的比值项可被解释为个体因果效应 β_{1i} 的加权平均。权重是 π_{1i}，它测量了工具变量影响第 i 个个体是否接受处理的程度。因此，TSLS 估计量是个体因果效应加权平均的一个一致性估计量，其中得到最大权重的个体是工具变量对他们最有影响的那些个体。

为了理解这一点，考虑两种情形：一种情形是 TSLS 估计量是平均因果效应的一致估计量，另一种情形不是。首先，假设所有个体有相同的因果效应，因此对于所有的 i 有 $\beta_{1i} = \beta_1$，那么，公式(11.11)中的最后一个表达式可简化为 $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \beta_1 E(\pi_{1i})/E(\pi_{1i}) = \beta_1$，所以该 TSLS 估计量就是因果效应 β_1 的一个一致估计量。也就是说，如果对于所有的个体其因果效应都是一样的，那么这个因果效应被 TSLS 一致地估计了，即使在工具变量对处理概率的影响中存在异质性。其次，假设存在异质因果效应，但工具变量等量地影响每个个体，因此有 $\pi_{1i} = \pi_1$。在此情形下，公式(11.11)中的最后一个表达式可简化为 $E(\beta_{1i}\pi_{1i})/E(\pi_{1i}) = \pi_1 E(\beta_{1i})/\pi_1 = E(\beta_{1i})$，该 TSLS 估计量就是平均因果效应 $E(\beta_{1i})$ 的一个一致估计量。因此，如果该工具变量对所有个体的影响相同，那么这个 TSLS 估计量就是平均因果效



① 原文方程中斜率系数为 π_{1i}，疑是印刷错误——译者注。

应的一致估计量，即使在处理效应中存在总体异质性也是如此。

再次，假设对一半的总体而言，Z_i 对处理决策没有影响（对它们而言，$\pi_{1i}=0$），而 Z_i 对另一半总体有一样的非零影响（对它们而言，π_{1i}是非零常数），那么，在这个工具变量影响处理决策的那一半总体中，TSLS 是平均因果效应的一致估计量。例如，假设一个工作培训项目的主体被给予一个随机的优先数 Z，它能影响（但不是决定）主体能否注册参加培训的资格。一些主体已具有培训项目所讲授的简历写作技能（对他们而言，$\beta_{1i}=0$），他们认为该培训项目是在浪费时间，不论他们的优先数是多少，都不会参加这个培训（因此，$\pi_{1i}=0$）。不过，对于需要该项目所讲授技能的主体而言（对他们而言，$\beta_{1i}>0$），优先数则起作用。在这个例子中，TSLS 估计了愿意参加该项目的那些主体的平均因果效应。对于愿意参加该项目的那些主体，他们的平均因果效应比全部总体的平均因果效应大，全部总体包括项目对他们有效的那些主体和项目对他们无效的那些主体。

一般地说，方程（11.11）中的最后那个比值项是个体因果效应β_{1i}的一个加权平均，用工具变量对处理的影响力来加权。如果工具变量的效应 π_{1i}独立于因果效应 β_{1i}，那么 $E(\beta_{1i}\pi_{1i})=E(\beta_{1i})E(\pi_{1i})$，因此 $E(\beta_{1i}\pi_{1i})/E(\pi_{1i})=E(\beta_{1i})$，进而 TSLS 是该因果效应的一个一致估计量。然而在实际中，由于是否接受处理（参与该项目）的个体决策依赖于他或她认为处理将会给他（她）们带来多少好处，因此，该工具变量的影响可能与因果效应有关。因此，在方程（11.11）的加权平均计算中，对工具变量在决定是否接受处理时起重要影响作用的那些个体的因果效应，将给予较大的权重，而对工具变量在决定是否接受处理时不起重要影响作用的那些个体的因果效应，将给予较轻的权重。

含义。上述的讨论有两个含义。首先，在通常 OLS 是一致的情况下，即当 $E(u_i|X_i)=0$ 时，该 OLS 估计量在总体存在异质因果效应时仍然是一致的，然而，由于不存在单个因果效应，因此该 OLS 估计量最好被解释为所研究总体的平均因果效应的一个一致估计量。其次，如果一个个体接受处理的决策依赖于该处理对那个个体的影响力，那么通常该 TSLS 估计量不是其平均因果效应的一致估计量。相反，TSLS 估计的是该因果效应的一个加权平均，其中受工具变量影响最大的那个个体因果效应得到的权重最大。这样会导致一种令人不安的情形，在此情形中，两个研究人员使用两个不同的但都是有效的工具变量（也即该工具变量既是相关的又是外生的），会得到对"这个"因果效应不同的估计值，甚至在大样本条件下也是如此。虽然这两个估计量通过它们各自方程（11.11）中的加权平均形式为因果效应的分布提供了一些洞察力，但是这两个估计量一般都不是其平均因果效应的一致估计量。①

例子：在心肌导管插入术研究中的应用。10.5 节和 11.5 节讨论了 McClellan，McNeil 和 Newhouse（1991）所做的心脏病人心肌导管插入术对死亡率影响的研究。作者使用了工具变量回归，用到心肌导管插入术医院的相对距离作为工具变量。根据他们的 TSLS 估计值，他们发现心肌导管插入术对健康结果几乎没有或没有影响。这个结果是令人惊讶的：像心肌导管插入术这样的治疗方法在被批准广泛使用之前要经历严格的临床实验。此外，心肌导管插入术允许外科医生执行医疗干预，该医疗干预在 10 年以前就应该要求做大的医疗手术，这样这些医疗干预会更安全，并且长期来看可能对病人的健康更有好处。这项经济计量



① 有不少很好的（但是高级的）关于总体异质性对项目评价估计量影响的研究。这些研究包括：Heckman，Lalonde 和 Smith（1999，第 7 节）所做的一项调查，以及 James heckman 在接受诺贝尔经济学奖时发表的演讲（Hecman（2001，第 7 节））。最近的参考文献以及 Angrist，Graddy 和 Imbens（2000）提供了随机效应模型（该模型将 β_{1i}看做为在个体间变化）的详细讨论和方程（11.11）中结论的更一般形式。

研究怎么没有发现心肌导管插入术的有益影响呢?

一种可能的答案是,在心肌导管插入术的处理效应中存在异质性。对一些病人而言,这是个有效的干预,但对其他病人,特别是对那些更健康的病人而言,这个疗法可能不太有效,或者在已知任何外科手术所面临的风险的条件下,这个疗法可能是根本无效的。因此,在心脏病人的这个总体中,该平均因果效应可能是(或大概是)正的。这样,该 IV 估计量测量了边际效应,而不是平均效应,其中,这里的边际效应是指,该疗法对于那些到医院的距离是他们是否接受处理的一个重要因素的病人的效应,但是,这些病人可能正好是身体状况相对较好的一些病人,对这些病人而言,从边际意义上讲,心肌导管插入术是个相对无效的疗法。如果是这样的话,McClellan,McNeil 和 Newhouse(1991)的 TSLS 估计量,测量了该疗法对那些边际病人(对他们而言,它是相对无效的)的效应,而不是对普通病人(对他们而言,它可能是有效的)的效应。

11.8 结论

在第 1 章中,我们根据一个理想的随机化控制实验的期望结果,定义了因果效应。如果一个理想的随机化控制实验是可获得的或是可实行的,那么该实验能够对所研究的因果效应提供一个强有力的证据,尽管随机化控制实验受到内部有效性和外部有效性的一些潜在的重要威胁。

虽然随机化控制实验存在诸多优势,但是它在经济学中仍然面临着一些困难,包括道德问题和成本问题。不过,实验方法的精髓可被应用于准实验。在准实验中,特殊的环境使它看上去"仿佛"随机化已经发生了。在准实验中,可使用差分再差分估计量来估计因果效应,这个因果效应可能被额外回归因子加以扩展。如果这个"仿佛"随机化只是部分地影响到处理,那么就可以使用工具变量回归。准实验的一个重要优势是,在数据中这个"仿佛"随机化的来源通常是明显的,因而可用具体的方法进行评价。准实验所面临的一个重要威胁是,有时这个"仿佛"随机化不是真正随机的,因此,这里的处理(或工具变量)与遗漏变量相关,这样相应得出的因果效应估计量是有偏的。

准实验在观测数据集和真正的随机化控制实验之间搭建了一座桥。本章中分析准实验所使用的经济计量方法是前面不同章节中所阐述的那些方法:OLS 法、面板数据估计法和工具变量回归。准实验与第 2 部分和第 3 部分中的研究应用的区别在于,解释这些方法的方式和所应用的数据集不同。准实验为经济计量学家提供了一种思考如何获得新数据集的方法,提供了一种如何思考工具变量的方法,提供了一种评价支撑 OLS 和工具变量估计的那些外生性假设合理性的方法。[①]

总结

1. 因果效应是根据一个理想的随机化控制实验定义的,因果效应可以用处理组和控制组的平均结果之差来估计。由于各种实际的原因,以人为主体的实验与理想的实验经常有



① Shadish,Cook 和 Campbell(2002)提供了一种在社会学和心理学中实验和准实验的综合处理方法。经济学中的实验例子包括负所得税实验(例子请见 www aspe. hhs. gov/hsp/sime-dime83)和兰德健康保险实验(the Rand health insurance experiment)(Newhouse(1983))。要了解更多的关于经济学中的准实验,见 Meyer(1995),Rosenzweig 与 wolpin(2000),以及 Angrist 与 Krueger(2001)。

偏离，一个重要原因是人们没有遵守实验协议。

2. 如果实际的处理水平 X_i 是随机的，那么可用实验结果对处理的回归来估计相应的处理效应，也可以将额外的预先处理的特征作为回归因子来改善回归的效率。如果分配的处理 Z_i 是随机的，但实际的处理 X_i 部分地是由个体选择决定的，那么相应的因果效应可使用以 Z_i 为工具变量的工具变量回归进行估计。

3. 在一项准实验中，法律或环境或自然事故的变化可被看做它们"仿佛"引致了对处理组和控制组的随机分配。如果实际的处理是"仿佛"随机的，那么其因果效应可用回归方法进行估计（也可能将额外预先处理的特征作为回归因子）；如果分配的处理是"仿佛"随机的，那么相应的因果效应可用工具变量回归进行估计。

4. 对一项准实验研究内部有效性的一个重要威胁是，这个"仿佛"随机化是否实际导致了外生性。由于存在行为反应，因此，就一个有效工具变量的要求条件来说，一个工具变量是由"仿佛"随机化所生成的，并不意味着它一定是外生的。

5. 当处理效应随个体变化时，如果实际的处理是被随机分配的或是被"仿佛"随机分配的，那么相应的 OLS 估计量就是平均因果效应的一个一致估计量。不过，这个工具变量估计量是个体处理效应的一个加权平均，其中最有影响的工具变量所对应的个体得到最大的权重。

重要术语

项目评估　因果效应　处理效应　差分估计量　局部依从性　损失　Hawthorne 效应　带有额外回归因子的差分估计量　条件均值独立性　差分再差分估计量　带有额外回归因子的差分再差分估计量　准实验　自然实验　重复截面数据　平均因果效应　平均处理效应

复习概念

11.1　一位研究人员想研究一种新肥料对农作物产量的效应。该研究人员计划进行一项实验，在这个实验中，给 100 块不同的一英亩面积的土地施不同数量的肥料。这将有四个处理水平：处理水平 1 是不施肥；处理水平 2 是只施厂家推荐肥料量的 50%；处理水平 3 是按 100% 推荐量施肥；处理水平 4 是按 150% 推荐量施肥。该研究人员计划对前 25 块土地实施处理水平 1，对第二个 25 块土地实施处理水平 2，依此类推。你能建议一个更好的分配处理水平的方法吗？为什么你的提议比该研究人员的方法好？

11.2　对一种降低胆固醇的新药进行临床实验。使用对病人的随机分配方法，给 500 个病人服用这种药，而给另外的 500 个病人服用安慰剂。你将如何估计这种药的处理效应？假设你有关于每个病人的体重、年龄以及性别的数据，你能用这些数据改善你的估计值吗？请给予解释。假设你有每个病人在进人该实验之前的胆固醇含量水平的数据，你能使用这些数据改善你的估计值吗？请给予解释。



11.3　研究 STAR 数据的研究人员报道了一些佚事证据，说家长向校长施加压力以将他们的孩子安置在小班。假设一些校长屈服于这种压力，将一些孩子转到了小班，这将会如何影响该项研究的内部有效性？假设你手头上有校长干预之前关于每个学生的最初随机分配的数据，你如何使用这个信息来修复这项研究的内部有效性？

11.4　请解释实验效应(像 Hawthorne 效应)在前面三个问题的每个实验中是否会很重要?

11.5　10.1 节中给出了一些学校受到地震毁坏这样一个假设的例子。请解释为什么这是一个准实验的例子,你如何使用这个事件所引起的班级规模的变化来估计班级规模对考试成绩的效应。

练习

*11.1　使用表 11—1 的结果,对每个年级计算下列结果:相对于常规班来说,小班处理效应的估计值、标准误,以及 95% 的置信区间(对于这个练习,不用考虑有助教的常规班的结果)。

11.2　对下列计算,使用表 11—2 第(4)列中的结果。考虑有这样两个班,它们与表 11—2 第(4)列中的回归因子有相同的值,这两个班记为 A 和 B,以下特征不同:

a. A 班是小班,而 B 班是常规班,构造平均考试成绩期望差异的 95% 的置信区间。

b. A 班有一个 5 年教龄的教师,而 B 班有一个 10 年教龄的教师,构造平均考试成绩期望差异的 95% 的置信区间。

c. A 班是一个小班,其教师有 5 年的教龄;而 B 班是一个常规班,但其教师有 10 年的教龄,构造平均考试成绩期望差异的 95% 的置信区间。(提示:在 STAR 中,教师是被随机地分配到不同类型的班级中的)

d. 为什么第(4)列中没有截距项?

11.3　考虑这样一项研究,该研究欲评估在大学生宿舍安装网络连接对大学生考试分数的影响。在一个大的宿舍楼中,有一半的寝室被随机地安装了高速宽带网络连接(处理组),并搜集了所有住宿学生最后一门课程的考试分数。下面哪个因素会形成对内部有效性的威胁? 为什么?

a. 在这个学年的中途,所有男运动员进入了大学生联谊会,并退出了该项研究(他们最终的考分没有被观测到)。

*b. 被分配到控制组的工程学专业的学生,他们建立了一个局域网络连接,这样他们既可以用较低的价格以集体方式付费,又可以个别地上网。

c. 处理组中艺术类专业的学生从来没有学会如何上网。

d. 处理组中的经济学专业,为那些控制组中的学生支付上网费。

11.4　假设一项随机化控制实验有 $T=2$ 期的面板数据,其中第一期的观测值($t=1$)是在实验之前取得的,而第二期的观测值($t=2$)是在处理后期取得的。假设该处理是一个二元变量,即如果第 i 个个体在处理组且 $t=2$,那么 $X_{it}=1$;否则,$X_{it}=0$。进一步假设,可以使用如下设定建立处理效应模型:

$$Y_{it}=\alpha_i+\beta_1 X_{it}+u_{it} \tag{11.12}$$

其中,α_i 是均值为 0、方差为 σ_α^2 的因个体不同而不同的效应(见方程(8.10)),而 u_{it} 是误差项,这里对于所有的 i,u_{it} 是同方差的,$\text{cov}(u_{i1},u_{i2})=0$,且 $\text{cov}(u_{it},\alpha_i)=0$。假设 $\hat{\beta}_1^{differences}$ 表示差分估计量,也即含有截距项的 Y_{i2} 对 X_{i2} 回归中的 OLS 估计量;再假设 $\hat{\beta}_1^{diffs\text{-}in\text{-}diffs}$ 表示差分再差分估计量,也即基于 $\Delta Y_i=Y_{i2}-Y_{i1}$ 对 $\Delta X_i=X_{i2}-X_{i1}$ 和截距的 OLS 回归的 β_1 估计量。

*a. 证明:$n\ \text{var}(\hat{\beta}_1^{differences})\rightarrow(\sigma_u^2+\sigma_\alpha^2)/\text{var}(X_{i2})$。(提示:使用附录 4.4 中 OLS 估计量



方差的同方差惟一的计算公式)

b. 证明：$n\ \mathrm{var}(\hat{\beta}_1^{diffs\ in\ diffs})\to 2\sigma_u^2/\mathrm{var}(X_{i2})$。(提示：注意 $\Delta X_i=X_{i2}$，为什么？)

c. 根据(a)和(b)中你的答案，基于纯粹的有效性考虑，在什么条件下你会选择差分再差分估计量而不选择差分估计量？

11.5 假设你有一项实验的 $T=2$ 期(因此 $t=1,2$)的面板数据，考虑这样一个面板数据回归模型，其中个体效应和时间效应是固定的，个体特征 W_i 不随时间的变化而变化(例如性别)。假设处理是二元变量，因此处理组中的个体在 $t=2$ 时，$X_{it}=1$；否则，$X_{it}=0$。考虑以下的总体回归模型：

$$Y_{it}=\alpha_i+\beta_1X_{it}+\beta_2(D_t\times W_i)+\beta_0D_t+v_{it} \tag{11.13}$$

其中，α_i 是个体固定效应；D_t 是个二元变量，如果 $t=2$，那么 $D_t=1$，如果 $t=1$，那么 $D_t=0$；$D_t\times W_i$ 是 D_t 和 W_i 之积；α 和 β 是未知系数。假定 $\Delta Y_i=Y_{i2}-Y_{i1}$，由公式(11.13)推导公式(11.15)(在单个回归因子 W 情况下，$r=1$)。

11.6 假设你拥有和练习 11.5 中的数据一样的数据(两个时期、n 个观测值的面板数据)，但忽略回归因子 W。考虑另一个可供选择的回归模型：

$$Y_{it}=\beta_0+\beta_1X_{it}+\beta_2G_i+\beta_3D_t+u_{it} \tag{11.14}$$

其中，如果个体在处理组，则 $G_i=1$；如果个体在控制组，则 $G_i=0$。证明：公式(11.14)中 β_1 的 OLS 估计量就是公式(11.13)中的差分再差分估计量。(提示：参见 6.3 节)

11.7 推导公式(11.9)中最后一个等式。(提示：使用协方差的定义，以及 β_{1i} 和 X_i 因实际处理 X_i 是随机的而是独立分布的这一事实)

附录 11.1 STAR 项目数据集

可公共接触的 STAR 项目数据集，包含了从 1985—1986 学年度到 1988—1989 学年度的 4 年实验期间的考试成绩、处理组以及学生和教师特征的数据集。本章中所分析的考试成绩数据是斯坦福达标考试的数学和阅读部分的成绩之和。表 11—2 中的二元变量“男生”表示学生是男生(=1)还是女生(=0)；二元变量“黑人”和“除黑人或白人因素外的种族因素”表示学生种族；二元变量“享有免费午餐资格”表示该学生在这个学年是否享有免费午餐计划的资格。教龄就是应用考试数据的某一年级学生所拥有的教师总教龄。该数据集还指明了在给定的年份里，学生所上的是哪一所学校，这使得构造特定学校的二元指示变量成为可能。

附录 11.2 差分再差分估计量推广到多个时期[1]

当超过两个时期时，因果效应可用第 8 章的固定效应回归模型进行估计。

首先，考虑没有额外的回归因子“W”的情形，那么总体回归模型就是合并的时间与个体固定效应的回归模型(公式(8.18))：

$$Y_{it}=\beta_0+\beta_1X_{it}+\gamma_2D2_i+\cdots+\gamma_nDn_i+\delta_2B2_t+\cdots+\delta_TBT_t+v_{it} \tag{11.15}$$

其中，$i=1,\cdots,n$ 表示个体，$t=1,\cdots,T$ 表示测量的时期。如果第 i 个个体在日期 t 已接受处理，那么 $X_{it}=1$；否则，$X_{it}=0$。$D2_i$ 是表示第 i 个个体的二元变量(即如果 $i=2$，那么 $D2_i=1$；



[1] 本附录利用8 3 节和8 4 节的资料。

否则,$D2_t=0$)。$B2_t$ 是表示第 2 个时期的二元变量,其他二元变量的定义类似。v_{it} 是误差项,$\beta_0,\beta_1,\gamma_2,\cdots,\gamma_n,\delta_2,\cdots,\delta_T$ 是未知系数。引入个体效应(表示每个个体的二元变量)控制了可影响 Y 的那些不可观测的个体特性。引入时间效应(表示时期的二元变量)控制了可影响到实验结果的各个时期间的差异,不管该个体是在处理组还是在控制组。例如,在一项工作培训项目实验实施期间发生了经济衰退,当 $T=2$ 时,公式(11.15)中的时间与固定效应回归模型可简化为公式(11.14)中的差分再差分回归模型。公式(11.15)中 β_1 的估计方法已在 8.4 节中做了讨论。

测量预先处理的特征或测量不随时间变化的特征的额外回归因子(W),可以被合并到固定效应回归框架中。就如在公式(11.5)上下文中所讨论的,在带有额外回归因子的差分再差分设定中,回归因子 W 影响的是不同时期 Y 的变化,而不是其水平值。例如,不论一个个体是否参与了某个工作培训项目,该个体以前的教育水平都是一个可能会影响收入变化的可观测因素。因此,为了将公式(11.5)推广到多期,回归因子 W 与时间效应二元变量是相互作用的。为了方便起见,假设只有单个 W 回归因子,那么公式(11.5)的多期推广形式是:

$$Y_{it}=\beta_0+\beta_1X_{it}+\beta_2(B2_t\times W_i)+\cdots+\beta_T(BT_t\times W_i)+\gamma_2D2_i+\cdots+\gamma_nDn_i+\delta_2B2_t+\cdots+\delta_TBT_t+v_{it} \tag{11.16}$$

其中,回归因子 $B2_t\times W_i$ 是二元变量 $B2_t$ 和 W_i 之间的交互作用。当只有两个时期时,带有个体固定效应、时间效应、回归因子 W 以及与单个时间二元变量 $B2_t$ 有交互作用的 W 的总体回归模型,同公式(11.5)中的总体回归模型相同(见练习 11.5)。

多期面板数据也可被用于描绘随时间变化的因果效应,例如,工作培训项目对收入的效应随着时间的推移是持续存在还是消失了。这样做的方法,在第 13 章使用时间序列数据估计因果效应的相应内容中做了讨论。

附录 11.3 条件均值独立性

本附录讨论在 11.3 节所提到的条件均值独立性假设及其在对一个普通处理效应 β_1 的估计中的作用。本讨论集中于带有额外回归因子的差分估计量(公式(11.2)中的 $\hat{\beta}_1$),但是这里的思想可推广到带有额外回归因子的差分再差分估计量中。

所谓的**条件均值独立性**(conditional mean independence)假设就是指,公式(11.2)中的误差项 u_i 的条件均值可依赖于预先处理的特征 $W_{1i},\cdots,W_{ri}$,但不依赖于 X_i,具体表示就是:

$$E(u_i|X_i,W_{1i},\cdots,W_{ri})=\gamma_0+\gamma_1W_{1i}+\cdots+\gamma_rW_{ri} \tag{11.17}$$

在条件均值独立性假设下,u_i 中不可观测的特征可能和可观测的预先处理的特征(W)相关,但是给定 W 时,u_i 的条件均值并不依赖于这个处理。

如果 W_i 是一组完全的二元指示变量,那么公式(11.17)中的线性假设并不是约束性的。如果变量 W 是连续的,那么公式(11.17)中的线性条件期望可被解释为在适当重新定义 W 的情况下的一个非线性条件期望。例如,如在 6.2 节中所讨论的,公式(11.17)右边的额外项可能是一个最初连续变量 W 的多项式函数。

考虑公式(11.17)成立的三种情形是很有用的。第一,如果重要概念 5.4 中的第一个最小二乘假设成立,那么 $E(u_i|X_i,W_{1i},\cdots,W_{ri})=0$,因此公式(11.17)的条件得到满足,并且其条件期望等于零。第二,如果这里的处理 X_i 是被随机分配的,进而独立地分布于所有



个体的特征，那么不论这些特征是可观测的已被包含在回归中（变量 W），还是不可观测的包含在误差项中，公式(11.17)都成立。如果 X_i 独立分布于 u_i 和 W_i，那么给定 W_i 和 X_i 下 u_i 的条件分布并不依赖于 X_i，特别是那个条件分布的均值并不依赖于 X_i（即使它可能依赖于 W_i）。在工作培训项目那个例子中，如果处理是被随机地分配的，那么它将不会获得前置教育水平的效应，不论教育是一个内含回归因子还是误差项中的遗漏部分。第三，这里的处理 X_i 以 W_i 为条件被随机地分配。在此情形中，u_i 的均值不依赖于 X_i，因为给定 W_i，处理是被随机分配的。如果以 W_i 为条件，且 X_i 和 u_i 是独立的，那么给定 W_i 下 u_i 的条件分布就不依赖于 X_i，因此它的条件均值也不依赖于 X_i，即使它可能依赖于 W_i。如果 W_i 是一组指示变量，那么条件均值独立性意味着在每组内或在指示变量所定义的“块”内，X_i 是被随机分配的，但是这个分配概率会在不同的块之间变化。在个体所组成的块之内的随机分配，有时被称为块随机化（block randomization）。

在条件均值假设下，β_1 是处理效应。为了弄明白这一点，计算公式(11.2)两边的条件期望：

$$E(Y_i|X_i,W_{1i},\cdots,W_{ri}) = \beta_0+\beta_1X_i+\beta_2W_{1i}+\cdots+\beta_{r+1}W_{ri}+E(u_i|X_i,W_{1i},\cdots,W_{ri})$$
$$=\beta_0+\beta_1X_i+\beta_2W_{1i}+\cdots+\beta_{r+1}W_{ri}+\gamma_0+\gamma_1W_{1i}+\cdots+\gamma_rW_{ri} \tag{11.18}$$

其中，第二个等式是由条件均值独立性假设（公式(11.17)）导出来的。评价在 $X_i=1$ 处（处理组）和在 $X_i=0$ 处（控制组）公式(11.8)中的条件期望并相减，得到：

$$E(Y_i|X_i=1,W_{1i},\cdots,W_{ri})-E(Y_i|X_i=0,W_{1i},\cdots,W_{ri})=\beta_1 \tag{11.19}$$

公式(11.19)左边就是在给定特征 W 下的个体被随机地分配到处理组和控制组中的一项实验所定义的因果效应，而且该因果效应是实验结果的期望值。由于这个因果效应不依赖于 W，因此它也是总体中一个被随机选择的成员的因果效应。

当公式(11.17)（连同重要概念 5.4 中的第二个到第四个最小二乘假设）成立时，带有额外回归因子的差分估计量就是一致的。直观上理解，通过引入 W_i 作为回归因子，该差分估计量就控制了处理概率会依赖于 W_i 这一事实。在条件均值独立性假设下，$\hat{\beta}_1$ 是一致的这一数学论证涉及矩阵代数，留到练习 16.9 中再做讨论。

条件均值独立性还提供了一个解释带有可观测数据回归的框架。在这些回归中，一些系数（关于“控制”变量的）并没有因果解释，但其他一些系数却有因果解释，如表 5—2、表 6—3 和表 7—2 中的一样。

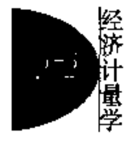
附录 11.4 当因果效应在个体间变化时的 IV 估计

本附录推导了公式(11.11)中的 TSLS 估计量的概率极限，其中，在处理效应和工具变量对处理接受的影响中存在总体异质性。具体来说，假设不包括带有异质效应的公式(11.8)和公式(11.10)成立这个条件，重要概念 10.4 中的 IV 回归假设成立。进一步假设：π_{1i}和β_{1i}独立分布于 u_i，v_i 和 Z_i，$E(u_i|Z_i)=E(v_i|Z_i)=0$，且 $E(\pi_{1i})\neq 0$。

由于$(X_i,Y_i,Z_i)(i=1,\cdots,n)$是独立同分布的，且有四阶矩，因此重要概念 2.6 中的大数定律适用。

$$\hat{\beta}_1^{TSLS}=s_{ZY}/s_{ZX}\xrightarrow{P}\sigma_{ZY}/\sigma_{ZX} \tag{11.20}$$

（见附录 3.3 和练习 15.2）。所以，这里的任务就是根据 π_{1i}和 β_{1i}的矩获得 σ_{ZY}和 σ_{ZX}的表达式。现已知 $\sigma_{ZX}=E[(Z_i-\mu_Z)(X_i-\mu_X)]=E[(Z_i-\mu_Z)X_i]$，将公式(11.10)代入这个 σ_{ZX}



第 4 部分

经济时间序列数据的回归分析

第 12 章

时间序列回归与预测导论

时间序列数据——即对单个实体在多个时点所搜集的数据——可用来回答那些单凭截面数据不足以回答的定量性问题。举一个这样的问题:一个变量 X 的变化对我们所感兴趣的另一个变量 Y 随时间变化的因果效应是什么? 换句话说,X 的变化对 Y 的动态因果效应是什么? 例如,一项要求乘客系安全带的法律,由于司机从一开始到后来对该法律有一个适应的过程,因此,随着司机对该法律的逐渐适应,该法律对交通死亡率的效应有多大? 另一个这样的问题是,你对某个变量在未来某个日期的最佳预测值是多少? 例如,你对下个月的通货膨胀率、利率或股票价格的最佳预测值是多少? 这两个问题,一个是关于动态因果效应的问题,另一个是关于经济预测的问题,它们都能用时间序列数据来解答。但是,时间序列数据存在一些固有的挑战,克服这些挑战需要一些新的技术。

第 12 章 ~ 第 14 章介绍了时间序列数据的经济计量分析技术,并将这些技术应用到预测和动态因果效应的估计问题中。第 12 章介绍了时间序列数据回归的基本概念和工具,并将它们应用于经济预测。第 13 章把第 12 章中所提出的概念和工具应用到用时间序列数据估计动态因果效应的问题中。第 14 章着手研究时间序列分析中一些更高级的问题,包括预测多元时间序列,对波动性随时间变化进行建模。

本章所研究的实证问题是预测通货膨胀率,即总体价格水平的百分比增加。虽然在某种意义上说,预测只是回归分析的一个应用,但是预测完全不同于因果效应的估计,这也是本书到目前为止的焦点问题。正如在 12.1 节中所讨论的,对预测有用的模型不一定需要有因果关系的解释:如果你看到行人拿着雨伞,你可能会预测要下雨,尽管拿雨伞并不会引起下雨。12.2 节介绍了时间序列分析的一些基本概念,并给出了经济时间序列数据的一些例子。12.3 节给出了回归因子是因变量的过去值的时间序列回归模型,这些“自回归”模型用通货膨胀的历史预测它的未来。通常,可通过增加额外的预测因子变量及这些变量的过去值或“滞后值”作为回归因子来改进以自回归为基础的预测,这些所谓的自回归分布滞后模

型在12.4节中介绍。例如,我们发现,在做通货膨胀预测时,除使用通货膨胀滞后值外,我们使用失业率的滞后值进行预测(也就是以经验的菲利浦斯曲线为基础的预测),可以改进自回归通货膨胀预测。一个实际问题是,在自回归模型和自回归分布滞后模型中包括多少个过去值才是合适的?12.5节介绍了解决这个问题的方法。

未来与过去是相似的,这是时间序列回归中的一个重要假设,这个假设有一个特定的称谓,即"平稳性"。时间序列变量可能不满足于不平稳性,其原因是多种多样的,但是在经济时间序列数据的回归分析中有两个原因非常重要:(1)序列可能具有持续的、长期的运动,即该序列含有趋势;(2)总体回归随时间的推移可能是不平稳的,即总体回归可能含有突变。这些与平稳性的背离危害了基于时间序列回归之上的预测和推断。幸运的是,存在用来发现趋势和突变的统计方法,并且一旦趋势和突变被发现,还存在调整模型设定的统计方法。这些方法在12.6节和12.7节中介绍。

12.1 使用回归模型进行预测

第4章我们从教育主管所考虑的一个问题开始,该教育主管想知道如果她削减其所在学区的班级规模,那么考试成绩将会提高多少,也就是说,该教育主管想知道班级规模的变化对考试成绩的因果效应。因此,第2部分和第3部分重点讨论了根据观测数据使用回归分析方法估计因果效应的问题。

现在考虑一个不同的问题,一位家长搬到一个大城市并部分地依据当地的学区系统选择了该地区内的一个城镇。这位家长想知道,不同学区在标准化考试中的表现有多大的区别。不过,假设无法得到考试成绩数据(可能数据是机密的),但可获得班级规模的数据,那么,这位家长必须根据有限的信息来猜测不同学区在标准化考试的中表现如何不同。也就是说,这位家长的问题是,根据与考试成绩相关的信息(如班级规模)来预测某一给定学区内的平均考试成绩。

从概念上说,这位教育主管的问题和这位家长的问题完全不同。多元回归是解决这两个问题的一个有力工具,但由于问题性质不同,因此用来评价一个特定回归模型合适性的标准也不同。为了得到那位教育主管想要的因果效应的可靠的估计值,我们必须要顾虑在第7章中所讨论的问题:遗漏变量偏差问题、选择问题、联立因果关系问题等等。相反,要得到那位家长想要得到的可靠的预测值,重要的是所估计的回归要有较强的解释能力,即所估计的回归系数要精确,而且该回归模型应该是平稳的,也就是说,根据一组数据集所估计的回归能够可靠地被用于其他数据集的预测。

例如,回想一下第4章介绍的考试成绩对学生—教师比(STR)的回归:

$$\widehat{TestScore} = 698.9 - 2.28 \times STR \tag{12.1}$$

我们得出的结论是,这个回归对回答该教育主管所关心的问题是没有用的:斜率的OLS估计量是有偏的,因为存在遗漏变量,诸如学生集体的构成和他们其他的校外学习机会。该教育主管不能改变该地区的平均收入水平或不说英语的人数比例,这两个变量都会影响考试成绩。由于这些变量也与班级规模相关,因此存在遗漏变量偏差。因而,考试成绩对学生—教师比的回归生成一个学生—教师比变化对考试成绩影响的有偏估计量,公式(12.1)不能够回答那个教育主管的问题。



不过,公式(12.1)可能对那位试图选择一个学区的家长是有用的。诚然,班级规模并不是考试成绩的惟一决定因素,但从家长的角度来看,重要的是它是否是考试成绩的一个可

靠的预测因子。对预测考试成绩感兴趣的那位家长并不关心公式(12.1)中的系数是否估计了班级规模对考试成绩的因果效应。更确切地说,家长只是希望用该回归来解释不同学区考试成绩中的大部分变化,并希望回归结果是平稳的,也就是说,家长只是想将回归的结果应用到他们正考虑迁居的地区。尽管遗漏变量偏差使得公式(12.1)对回答因果关系问题是无用的,但它仍然可以用于预测。

本章中的应用不同于考试成绩与班级规模之间关系的预测问题,因为本章关注于使用时间序列数据预测未来事件,而时间序列预测在概念上类似于那位家长的问题:任务是使用某些变量的已知值(价格通货膨胀率的当期值和过去值,而不是班级规模)去预测另一个变量的值(未来通货膨胀率,而不是考试成绩)。像那位家长所关心的问题一样,即使回归模型的系数没有因果解释能力,该回归模型也能生成可靠的预测值。在第13章,我们回到那位教育主管所面对的问题,并讨论了如何使用时间序列变量估计因果效应的问题。

12.2 时间序列数据和序列相关知识介绍

本节介绍时间序列经济计量学中出现的一些基本概念和术语。分析任何时间序列数据的一个好的出发点都是绘制数据图形,因此我们就从绘制时间序列图形开始。

12.2.1 美国的通货膨胀率和失业率

图12—1(a)绘制了从1960年到1999年美国通货膨胀率的图形——由消费者价格指数(CPI)所测度的美国价格水平的年度百分比变化(在附录12.1中描述了该数据)。通货膨胀率在20世纪60年代是低的,在整个70年代一直上升,并在1980年第一季度(即1980年1月份、2月份、3月份)达到战后最高点15.5%,而后一直下降到90年代末期的3%以下。从图12—1(a)中可以看出,通货膨胀率从一个季度到下一个季度可能波动一个或多个百分点。

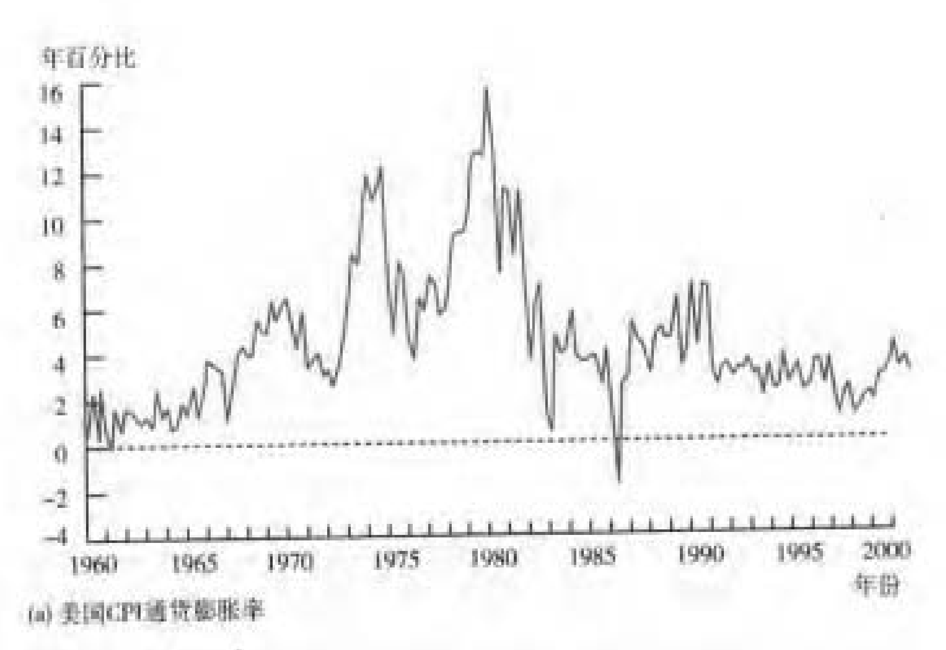
图12—1(b)绘制了美国失业率的图形(失业率是指没有工作的劳动力人口占总劳动力人口的比重,它是由美国的当前人口调查提供的(见附录3.1))。失业率的变化主要与美国的经济周期相联系。例如,在1960—1961年、1970年、1974—1975年的衰退期间,在1980年和1981—1982年的双衰退期间,以及在1990—1991年的衰退期间,失业率都上升了,由图12—1(b)中的阴影插条来表示。

12.2.2 滞后,一阶差分,对数和增长率

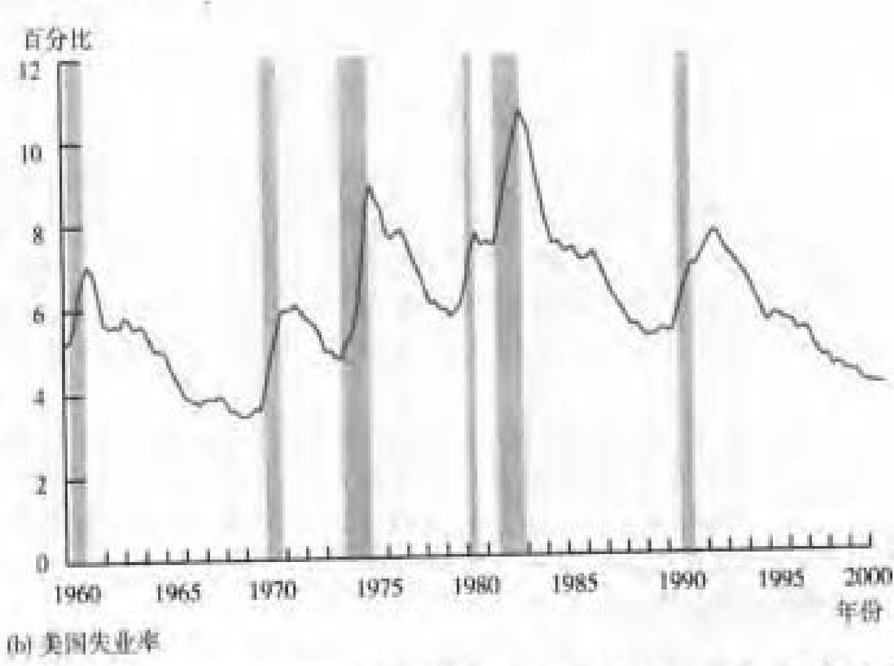
时间序列变量Y在t时期的观测值表示为Y_t,观测值的总数表示为T。观测值之间的时间间隔,即第t期的观测值和第$t+1$期的观测值之间的时期,通常是某些时间单位,比如说周、月、季(3个月度单位)或年。例如,本章所研究的通货膨胀数据是季度的数据,所以时间单位(一个"时期")是一年的四分之一。

对于Y的未来值和过去值,我们用一些专门的术语和符号来表示。Y的前一期的值叫做**一阶滞后值(first lagged value)**,或简称为**一阶滞后(first lag)**,记作Y_{t-1}。它的**第j阶滞后值(j^{th} lagged value)(或简称为第j阶滞后(j^{th} lag))**是其在j期以前的值,记为Y_{t-j}。同理,Y_{t+1}表示Y在将来一期的值。

Y的值在第$t-1$期和第t期之间的变化是Y_t-Y_{t-1},这个变化被称为变量Y_t的**一阶差分(first difference)**。在时间序列数据中,"Δ"被用来表示一阶差分,因此,$\Delta Y_t=Y_t-Y_{t-1}$。



(a) 美国CPI通货膨胀率



(b) 美国失业率

注：从1960年到1980年，美国价格水平通货膨胀率（图12—1(a)）向上移动，而后在20世纪80年代初期急剧下降。美国失业率（图12—1(b)）在衰退期间上升（阴影部分），而在扩张期间下降。

图12—1　1960—1999年美国通货膨胀与失业

分析经济时间序列数据经常是在计算了它们的对数或其对数的变化之后才进行的。这样做的一个原因是，许多经济序列如国内生产总值（GDP），都会表现出近似于指数形式的增长，也就是说，在一个较长的时期内，平均来说该序列倾向于每年以一定的百分比增长，如果是这样的话，该序列的对数就近似线性地增长。另一个原因是，许多经济时间序列的标准差与它们的水平值近似成比例，也就是说，其标准差可以被很好地表示为该对应序列水平值的百分比，如果是这样的话，那么该序列对数的标准差近似于常数。对于以上两种情况中的任何一种情况，进行序列变换都是有用的，它使得变换后的序列的变化与原始序列成比例地

（或百分比）变化，这可以通过对原始序列取对数得到。[①]

重要概念 12.1

滞后、一阶差分、对数和增长率

■一个时间序列 Y_t 的一阶滞后是 Y_{t-1}；它的 j 阶滞后是 Y_{t-j}。

■一个序列的一阶差分 ΔY_t，是该序列在第 $t-1$ 期和第 t 期之间的变化，即 $\Delta Y_t = Y_t - Y_{t-1}$。

■序列 Y_t 的对数的一阶差分是 $\Delta \ln(Y_t) = \ln(Y_t) - \ln(Y_{t-1})$。

■一个时间序列 Y_t 在第 $t-1$ 期和第 t 期之间的百分比变化约为 $100\Delta\ln(Y_t)$，这里当百分比变化很小时，近似程度最精确。

重要概念 12.1 总结了滞后、一阶差分、对数和增长率。

表 12—1 例示了美国通货膨胀率数据的滞后、变化和百分比变化。第一列表示日期或时期，其中 1999 年第一季度记为 1999：Ⅰ，1999 年第二季度记为 1999：Ⅱ，依此类推。第二列表示该季度 CPI 的值，第三列表示通货膨胀率。例如，从 1999 年的第一季度到第二季度，该指数从 164.9 增加到 166.0，百分比增加的幅度为 $100 \times (166.03 - 164.87)/164.87 = 0.704\%$，这就是前后两个季度间的百分比增加。习惯上，通货膨胀率是以年度为基准报告的（许多其他的宏观经济时间序列的增长率也是如此），以年度为基准的通货膨胀率说明，如果该序列会持续以同样的比率增加，那么价格在一年之中将会上升的百分比。因为一年有四个季度，所以在 1999：Ⅱ季度转换为年度的通货膨胀率为 $0.704 \times 4 = 2.82$，或四舍五入之后为每年 2.8%。



表 12—1　　　　1999 年和 2000 年第一季度的美国通货膨胀

季度	美国 CPI	年通货膨胀率 (Inf_t)	一阶滞后 (Inf_{t-1})	通货膨胀变化 (ΔInf_t)
1999: Ⅰ	164.87	1.6	2.0	-0.4
1999: Ⅱ	166.03	2.8	1.6	1.2
1999: Ⅲ	167.20	2.8	2.8	0.0
1999: Ⅳ	168.53	3.2	2.8	0.4
2000: Ⅰ	170.27	4.1	3.2	0.9

注：年度化的通货膨胀率是从前一季度到当前季度 CPI 的百分比变化乘以 4。通货膨胀的一阶滞后是它在前一季度的值，而通货膨胀的变化是当前通货膨胀率减去它的一阶滞后值。所有的数字都被四舍五入到一位小数。

这个百分比变化也可以用重要概念 12.1 中的对数差分近似公式来计算。从 1999：Ⅰ

① 回想一下 6.2 节，一个变量的对数的变化约等于该变量的比例变化，即 $\ln(X+a) - \ln(X) \approx a/X$，这里当 a/X 很小时近似效果最好。现以 Y_{t-1} 代替 X，以 ΔY_t 代替 a，并注意 $Y_t = Y_{t-1} + \Delta Y_t$。这意味着在第 $t-1$ 期和第 t 期之间序列 Y_t 的比例变化约为 $\ln(Y_t) - \ln(Y_{t-1}) = \ln(Y_{t-1} + \Delta Y_t) - \ln(Y_{t-1}) \approx \Delta Y_t / Y_{t-1}$。表达式 $\ln(Y_t) - \ln(Y_{t-1})$ 是 $\ln(Y_t)$ 的一阶差分，即 $\Delta\ln(Y_t)$。因此，$\Delta\ln(Y_t) \approx \Delta Y_t / Y_{t-1}$。百分比变化是分数变化的 100 倍，因此，序列 Y_t 的百分比变化约为 $100\Delta\ln(Y_t)$。

乍看起来，通货膨胀水平值之间强正相关，而通货膨胀的变化之间呈负相关，这好像是矛盾的，但是，这两个自相关测量的是不同的事情。通货膨胀的强正自相关反映了图12—1中显示的通货膨胀的长期趋势：在1965年第一季度，通货膨胀是低的，而在第二季度也是如此；在1981年第一季度，通货膨胀是高的，而在第二季度也是如此。相反，通货膨胀变化的负自相关意味着，平均来看，某个季度通货膨胀的上升往往与下一季度通货膨胀的下降相联系。

表12—2　**美国通货膨胀率及其变化的前四阶样本自相关系数：1960：Ⅰ—1999：Ⅳ**

	自相关系数	
滞后阶数	通货膨胀率(Inf_t)	通货膨胀率变化(ΔInf_t)
1	0.85	-0.24
2	0.77	-0.27
3	0.77	0.32
4	0.68	-0.06

12.2.4　经济时间序列的其他例子

经济时间序列大不相同。图12—2绘制了四个不同的经济时间序列的例子：美国联邦基金利率、美元对英镑的汇率、日本实际国内生产总值的对数和标准普尔500(S&P500)股票市场指数日收益率。

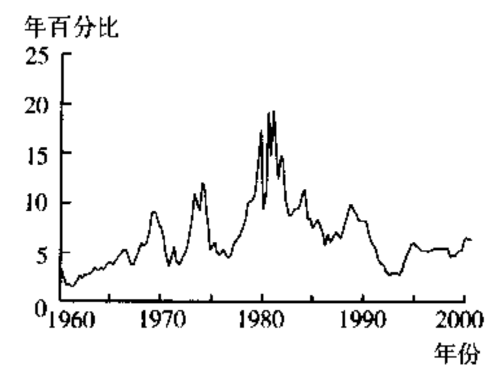
美国联邦基金利率(见图12—2(a))是银行之间相互支付的隔夜借款利率。这个利率是很重要的，因为它受美国联邦储备委员会控制，而且是美联储的主要货币政策工具。如果将联邦基金利率图形与图12—1中的失业率和通货膨胀率相比较，你会发现联邦基金利率的急剧上升经常与随后的衰退相联系。

美元/英镑汇率(见图12—1(b))是1英镑(£)的美元价格。在1972年以前，发达国家实行固定汇率制度，被称为“布雷顿森林”体系，在该体系下，政府的工作是防止汇率波动。在1972年，通货膨胀的压力导致了“布雷顿森林”体系的崩溃，随后，一些主要的货币允许“浮动”，也就是说，它们的价值由外汇市场上货币的供求来决定。在1972年以前，汇率几乎是常数，除1968年惟一一次贬值外，当时英镑相对美元的官方报价跌至2.40美元。自1972年以来，汇率已在一个非常广的范围内波动。

日本实际季度GDP(见图12—2(c))是经通货膨胀调整后，在一个季度内日本所生产的商品和服务的价值总和。GDP是总体经济活动的最广泛的测量指标。图12—2(c)绘制了该序列的对数，这个序列的变化可被解释为(小数的)增长率。在20世纪60年代和70年代初期，日本实际GDP增长迅速，但这种增长在70年代后期和80年代却缓慢下来。90年代期间，增长进一步减缓，从1990—1999年，年均增长率仅为1.5%。

纽约股票交易所(NYSE)股票价格指数日收益率(见图12—2(d))是NYSE综合市场指数从一个交易日到下一交易日的百分比变化，该指数是在纽约股票交易所交易的所有公司股票价格的一个最宽泛的指数。图12—2(d)绘制了从1990年1月2日到1998年12月31日(共1 771个观测值)的日收益率图形。不像图12—2中的其他序列，在这些日收益率序列中几乎不存在序列相关，如果存在，那么你就可以使用过去的日收益率来预测这些收益率，并且能够通过在预测市场开始上涨时买入而在预测市场开始下跌时卖出来赚钱。尽管收益率本身在实质上是不可预测的，但通过对图12—2(d)的检查，可以揭示收益率波动的

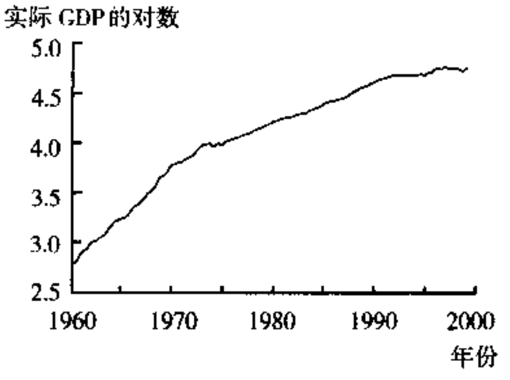
某些模式。例如,收益率的标准差在1991年和1998年相对较大,而在1995年相对较小。这种"波动集聚现象"在许多金融时间序列中被发现,对这种特殊类型异方差序列进行建模的经济计量模型在14.5节中进行了研究。



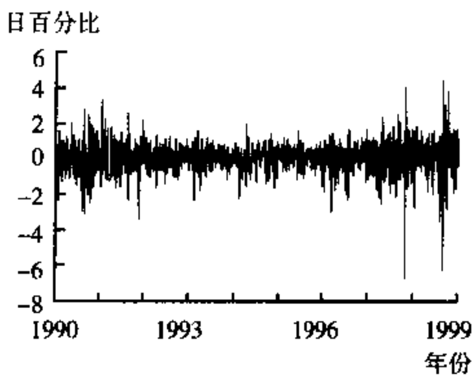
(a)联邦基金利率



(b)美元/英镑的汇率



(c)日本实际GDP的对数



(d) NYSE综合股票指数日价值的百分比变化

注:这四个时间序列具有明显不同的图形。联邦基金利率(见图12—2(a))有一个与价格水平通货膨胀类似的图形。美元和英镑之间的汇率(见图12—2(b))显示了在1972年固定汇率的布雷顿森林体系崩溃之后的离散变化。日本实际GDP的对数(见图12—2(c))表现出相对平滑的增长,尽管增长率在20世纪70年代是下降的,并且在90年代再一次下降。NYSE股票价格指数日收益率(见图12—2(d))在本质上是不可预测的,但它的方差是变化的,这个序列表现出"波动集聚"现象。

图12—2 四个经济时间序列

12.3 自回归

明年的价格水平通货膨胀率(即总体价格水平的百分比增加)将是多少?华尔街的投资者在决定对债券投资支付多少时要依靠通货膨胀的预测。像美国联邦储备银行这样的中央银行的经济学家在制定货币政策时要使用通货膨胀预测,企业在预测产品销售时要使用通货膨胀预测,地方政府在制定来年财政预算时也要使用通货膨胀预测。在本节中,我们考虑使用自回归(autoregression)所做的预测,自回归是一个将时间序列变量和它的过去值联系在一起的回归模型。

12.3.1 一阶自回归模型

如果你想要预测一个时间序列的未来值,一个好的出发点是考察它刚刚过去的邻近的

值。例如,如果你要预测这个季度到下个季度通货膨胀的变化,那么你可以看看上个季度通货膨胀是上升了还是下降了。利用前一个季度通货膨胀的变化 ΔInf_{t-1} 来预测本季度通货膨胀的变化 ΔInf_t,一个系统的方法就是估计 ΔInf_t 对 ΔInf_{t-1} 的 OLS 回归。利用从 1962—1999 年的数据进行估计,所得的回归方程是:

$$\widehat{\Delta Inf_t} = \underset{(0.14)}{0.02} - \underset{(0.106)}{0.211}\,\Delta Inf_{t-1} \tag{12.7}$$

其中,和通常一样,标准误在估计系数下面的括号中给出,$\widehat{\Delta Inf_t}$ 是以所估计的回归线为基础的 ΔInf_t 的预测值。公式(12.7)中的模型被称为一阶自回归,之所以被称为自回归,是因为它是序列对它自身的滞后值 ΔInf_{t-1} 的回归;而之所以被称为一阶,则是因为只使用一个滞后值作为回归因子。公式(12.7)中的系数是负的,因此一个季度通货膨胀率的上升往往伴随着下一个季度通货膨胀率的下降。

一阶自回归被简记为 AR(1),其中"1"表明它是一阶的。序列 Y_t 的总体 AR(1)模型是:

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t \tag{12.8}$$

其中,u_t 是误差项。

预测与预测误差。假设你有关于 Y 的历史数据,并且你想要预测它的未来值。如果 Y_t 满足公式(12.8)中的 AR(1)模型且 β_0 和 β_1 是已知的,那么基于 Y_{t-1} 的 Y_t 的预测值为 $\beta_0 + \beta_1 Y_{t-1}$。

在实际中,β_0 和 β_1 是未知的,因此要预测必须首先对 β_0 和 β_1 进行估计。我们将使用 OLS 估计量 $\hat{\beta}_0$ 和 $\hat{\beta}_1$,它们是使用历史数据构造出来的。一般地说,利用直到 $t-1$ 期的数据所估计的模型,$\hat{Y}_{t|t-1}$ 表示在直到 $t-1$ 期的信息基础上 Y_t 的预测值。因此,以公式(12.8)中的 AR(1)模型为基础的预测是:

$$\hat{Y}_{t|t-1} = \hat{\beta}_0 + \hat{\beta}_1 Y_{t-1} \tag{12.9}$$

其中,$\hat{\beta}_0$ 和 $\hat{\beta}_1$ 是利用直到 $t-1$ 期的历史数据估计出来的。

预测误差(forecast error)是预测所犯的错误,它是实际发生的 Y_t 的值与根据 Y_{t-1} 做出的 Y_t 的预测值之差。

$$\text{预测误差} = Y_t - \hat{Y}_{t|t-1} \tag{12.10}$$

预测与预测值。这里所说的预测并不是一个 OLS 预测值,这里所说的预测误差也并不是指 OLS 残差。OLS 预测值是为估计回归的那些在样本中的观测值而计算的。相反,预测值是为用于估计回归的数据集以外的某些时期而做的,因此,所预测的因变量的实际值的数据不在用来估计该回归的样本数据中。同理,OLS 残差是 Y 的实际值与它在样本内观测值的预测值之差,而预测误差是指 Y 的未来值(它不包含在估计样本中)与该未来值的预测值之差。换句话说,预测和预测误差(forecasts and forecast errors)从属于"样本外"观测值,而预测值和残差(predicted values and residuals)从属于"样本内"观测值。

均方根预测误差。均方根预测误差(root mean squared forecast error,简写为 RMSFE)是预测误差大小的一个测量,也就是说,它是使用某一预测模型所犯的错误大小的一个测量指标。RMSFE 是均方预测误差的平方根。

$$\text{RMSFE} = \sqrt{E[(Y_t - \hat{Y}_{t|t-1})^2]} \tag{12.11}$$

RMSFE 有两个误差来源:一个是因 u_t 的未来值是未知时所引起的误差,一个是在估计

系数β_0和β_1时所产生的误差。如果第一个来源的误差远大于第二个(如果样本容量很大时就可能出现这种结果),那么 RMSFE 近似等于总体回归(见公式(12.8))中误差项u_t的标准差$\sqrt{\mathrm{var}(u_t)}$。反过来,u_t的标准差是用回归(*SER*,见5.10节)的标准误估计出来的。因此,如果由估计回归系数所引起的不确定性小得足以被忽略,那么 RMSFE 就可以由回归的标准误差来估计。12.4节中研究了包含两个预测误差来源的 RMSFE 的估计。

在通货膨胀案例中的应用　根据公式(12.7)中所估计的 AR(1)模型(它是用直到1999:Ⅳ的数据来估计的),预测者在1999:Ⅳ就已应做出的2000年第一季度(2000:Ⅰ)的通货膨胀预测值是多少?由表12—1可知,1999:Ⅳ的通货膨胀率为3.2%(因此,$Inf_{1999:IV}=3.2\%$),比1999:Ⅲ增加0.4个百分点(因此,$\Delta Inf_{1999:IV}=0.4$)。将这些值代入公式(12.7),从1999:Ⅳ到2000:Ⅰ通货膨胀变化的预测值为$\widehat{\Delta Inf}_{2000:I}=0.02-0.211\times\Delta Inf_{1999:IV}=0.02-0.211\times0.4=-0.06\approx-0.1$(四舍五入到十位)。通货膨胀率的预测值为通货膨胀率的过去值加上预测的变化值:

$$\widehat{Inf}_{t|t-1}=Inf_{t-1}+\widehat{\Delta Inf}_{t|t-1} \tag{12.12}$$

因为$Inf_{1999:IV}=3.2\%$,从1999:Ⅳ到2000:Ⅰ通货膨胀预测的变化值是-0.1,所以2000:Ⅰ的通货膨胀率的预测值是$\widehat{Inf}_{2000:I}=Inf_{1999:IV}+\widehat{\Delta Inf}_{2000:I}=3.2\%-0.1\%=3.1\%$。因此,AR(1)模型预测出该通货膨胀将由1999:Ⅳ的3.2%稍微下降到2000:Ⅰ的3.1%。

这个 AR(1)预测值的精确度如何?由表12—1可知,2000:Ⅰ通货膨胀的实际值为4.1%,这样 AR(1)预测值整整低了1个百分点,即预测误差是1.0%。公式(12.7)中的 AR(1)模型的$\bar{R}^2$仅为0.04,所以滞后的通货膨胀变化只解释了用来拟合自回归样本中通货膨胀变化的非常小的部分。这个低的$\bar{R}^2$值与用公式(12.7)所生成的那个较差的2000:Ⅰ通货膨胀预测值是一致的。更一般地说,低的$\bar{R}^2$值表明,这个 AR(1)模型只能预测通货膨胀变化中的一小部分变化。

公式(12.7)的回归标准误为1.67。如果忽略由系数估计所引起的不确定性,那么对以公式(12.7)为基础的预测值,RMSFE 的估计值为1.67个百分点。

12.3.2　p阶自回归模型

AR(1)模型使用Y_{t-1}来预测Y_t,但是这样做忽略了在更远的过去所潜在的有用的信息。综合考虑这些信息的一种方法是,在 AR(1)模型中包含额外的滞后项,这就得到了p阶自回归模型,或称 AR(p)模型。

p阶自回归模型(**p^{th} order autoregressive model**,或称 **AR(p)模型**)将Y_t表示为它的p个滞后值的线性函数,即在 AR(p)模型中,回归因子是$Y_{t-1},Y_{t-2},\cdots,Y_{t-p}$和一个截距项。包含在 AR($p$)模型中的滞后数$p$,被称为自回归的滞后阶数或滞后长度。

例如,通货膨胀变化的 AR(4)模型用通货膨胀变化的四阶滞后作为回归因子。用 OLS 对1962—1999年的数据进行估计,所得的 AR(4)模型为:

$$\underset{}{\widehat{\Delta Inf_t}}=\underset{(0.12)}{0.02}-\underset{(0.10)}{0.21}\,\Delta Inf_{t-1}-\underset{(0.09)}{0.32}\,\Delta Inf_{t-2}+\underset{(0.09)}{0.19}\,\Delta Inf_{t-3}-\underset{(0.10)}{0.04}\,\Delta Inf_{t-4} \tag{12.13}$$



公式(12.13)中的最后三个额外滞后项的系数在5%的显著性水平下联合地显著异于0:F统计量是6.43(p值<0.001)。这反映在$\bar{R}^2$值由公式(12.7)中 AR(1)模型的0.04提高到 AR(4)模型的0.21上。同理,公式(12.13)中 AR(4)模型的 *SER* 为1.53,比 AR(1)模型的 *SER* 值1.67有所改进。

AR(p)模型在重要概念 12.3 中总结。

重要概念 12.3

自回归

p 阶自回归模型将 Y_t 表示为它的 p 个滞后值的线性函数：

$$Y_t=\beta_0+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+\cdots+\beta_p Y_{t-p}+u_t \tag{12.14}$$

其中，$E(u_t|Y_{t-1},Y_{t-2},\cdots)=0$。滞后数 p 被称为自回归的滞后阶数或滞后长度。

AR(p)模型中预测和误差项的性质。给定 Y_t 的过去值，u_t 的条件期望为 0（即 $E(u_t|Y_t,Y_{t-1},\cdots,Y_{t-p})=0$）这个假设具有两个重要的含义。

第一个含义是，基于 Y_t 的全部历史值所做出的它的最佳预测只依赖于最近的 p 个过去值。具体地讲，令 $Y_{t|t-1}=E(Y_t|Y_{t-1},Y_{t-2},\cdots)$ 为给定 Y_t 的所有历史值条件下 Y_t 的条件均值，那么在基于 Y 的历史值的所有预测中，$Y_{t|t-1}$ 具有最小的 RMSFE（见练习 12.5）。如果 Y_t 满足 AR(p)，那么它的条件均值为：

$$Y_{t|t-1}=\beta_0+\beta_1 Y_{t-1}+\beta_2 Y_{t-2}+\cdots+\beta_p Y_{t-p} \tag{12.15}$$

它是从公式（12.14）中的 AR(p)模型和 $E(u_t|Y_{t-1},Y_{t-2},\cdots)=0$ 的假设中得出的。实际上，系数 $\beta_0,\beta_1,\cdots,\beta_p$ 是未知的，因此，由 AR(p)得出的实际预测值需要使用带有估计系数的公式（12.15）。

第二个含义是，误差项 u_t 是序列不相关的，这是由公式（2.25）中得出的结论（见练习 12.5）。

在通货膨胀案例中的应用。根据公式（12.13）中通货膨胀的 AR(4)模型，使用直到 1999：Ⅳ的数据，2000：Ⅰ通货膨胀的预测值是多少？为了计算这个预测值，将 1999 年四个季度的每一个季度的通货膨胀变化值代入公式（12.13）中：$\widehat{\Delta Inf}_{2000\,\mathrm{I}|1999\,\mathrm{IV}}=0.02-0.21\Delta Inf_{1999\,\mathrm{IV}}-0.32\Delta Inf_{1999\,\mathrm{III}}+0.19\Delta Inf_{1999\,\mathrm{II}}-0.04\Delta Inf_{1999\,\mathrm{I}}=0.02-0.21\times 0.4-0.32\times 0.0+0.19\times 1.1-0.04\times(-0.4)\approx 0.2$，这里 1999 年通货膨胀变化值取自表 12—1 的最后一列。

相应的 2000：Ⅰ的通货膨胀的预测值是 1999：Ⅳ的通货膨胀值加上预测的变化值，即 3.2% +0.2% =3.4%。预测误差为实际值 4.1% 减去预测值，即 4.1% −3.4% =0.7%，略小于 AR(1)的预测误差 1.0%。

12.4 含有额外预测因子的时间序列回归与自回归分布滞后模型

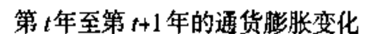
经济理论常常建议，其他一些变量在预测我们所感兴趣的变量时也会有所帮助。这些其他的变量，或称预测因子，可被添加到一个自回归中生成一个含有多个预测因子的时间序列回归模型。当其他的变量及其滞后项被添加到一个自回归中时，其结果就是一个自回归分布滞后模型。

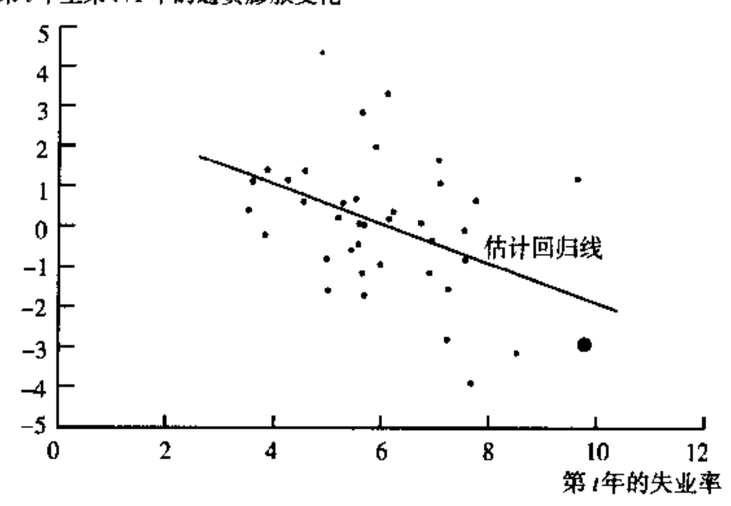
12.4.1 利用过去的失业率预测通货膨胀率的变化

高的失业率值倾向于与未来通货膨胀率的下降有联系。这种负的相关关系，即著名的短期菲利浦斯曲线，在散点图 12—3 中看得很清楚，图中价格通货膨胀率的年度变化对应于



上一年度的失业率。例如，1982 年平均失业率为 9.7%，在下一年通货膨胀率下降 2.9%。大体来说，图 12—3 中的相关系数为 -0.40。





注：1982 年，美国失业率为 9.7%，且 1983 年的通货膨胀率下降 2.9%（最大的点）。一般地说，t 年的高失业率值倾向于紧接着在下一年即 t+1 年有价格水平通货膨胀率的下降，且相关系数为 -0.40。

图 12—3　t 与 t+1 年之间通货膨胀变化和 t 年失业率的散点图

图 12—3 中的散点图表明，失业率的过去值可能包含了关于通货膨胀未来进程的信息，而这些信息并没有包含在通货膨胀的过去变化中。这个推测通过扩展公式（12.13）中的 AR(4) 模型把失业率的一阶滞后包含进来就很容易检验：

$$\widehat{\Delta Inf_t} = \underset{(0.55)}{1.42} - \underset{(0.09)}{0.26}\,\Delta Inf_{t-1} - \underset{(0.10)}{0.40}\,\Delta Inf_{t-2} + \underset{(0.08)}{0.11}\,\Delta Inf_{t-3} - \underset{(0.10)}{0.09}\,\Delta Inf_{t-4} - \underset{(0.10)}{0.23}\,Unemp_{t-1} \quad (12.16)$$

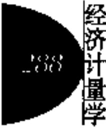
$Unemp_{t-1}$ 的 t 统计量为 -2.33，因此，这个项在 5% 的水平下是显著的。这个回归的 $\bar{R}^2$ 为 0.22，比 AR(4) 的 $\bar{R}^2$0.21 稍有改善。

将 1999 年通货膨胀的变化值连同 1999：Ⅳ 的失业率的值（4.1%）一起代人公式（12.16）中，便可得到 2000：Ⅰ 通货膨胀变化的预测值，所得的预测值为 $\widehat{\Delta Inf}_{2000:\mathrm{I}|1999:\mathrm{IV}} = 0.5$。因此，2000：Ⅰ 通货膨胀的预测值是 3.2% +0.5% =3.7%，预测误差是 0.4%。这个预测值比 AR(4) 的预测值更接近于 2000：Ⅰ 的实际通货膨胀水平。

如果失业率的一阶滞后有助于预测通货膨胀，那么多阶滞后可能是更有帮助的。再增加三个更高阶的失业率滞后项，相应的估计结果如下：

$$\begin{aligned}\widehat{\Delta Inf_t} = {} & \underset{(0.47)}{1.32} - \underset{(0.09)}{0.36}\,\Delta Inf_{t-1} - \underset{(0.10)}{0.34}\,\Delta Inf_{t-2} + \underset{(0.08)}{0.07}\,\Delta Inf_{t-3} - \underset{(0.09)}{0.03}\,\Delta Inf_{t-4} \\ & - \underset{(0.47)}{2.68}\,Unemp_{t-1} + \underset{(0.89)}{3.43}\,Unemp_{t-2} - \underset{(0.89)}{1.04}\,Unemp_{t-3} + \underset{(0.44)}{0.07}\,Unemp_{t-4} \quad (12.17)\end{aligned}$$

检验二至四阶失业率滞后项的联合显著性的 F 统计量为 4.93（p 值 =0.003），因此它们是联合显著的。公式（12.17）中回归的 $\bar{R}^2$ 为 0.35，比公式（12.16）的 0.22 有实质性改善。所有失业率系数的 F 统计量是 8.51（p 值 <0.001），这表明了该模型比 12.3 节（公式（12.13））中的 AR(4) 模型在统计上有了显著的改善。公式（12.17）中回归的标准误为 1.37，比 AR(4) 中的 *SER*1.53 有了较大改善。



利用公式（12.17），从 1999：Ⅳ 到 2000：Ⅰ 通货膨胀变化的预测值可通过将相应变量的值代入到方程中进行计算。1999：Ⅰ 和 1999：Ⅱ 的失业率为 4.3%，1999：Ⅲ 的失业率为

4.2%,1999:Ⅳ的失业率为4.1%。根据公式(12.17),从1999:Ⅳ到2000:Ⅰ通货膨胀变化的预测值为:

$$\begin{aligned}\widehat{\Delta Inf}_{2000:\mathrm{I}|1999:\mathrm{IV}} &= 1.32-0.36\times0.4-0.34\times0.0+0.07\times1.1-0.03\times(-0.4)\\ &\quad -2.68\times4.1+3.43\times4.2-1.04\times4.3+0.07\times4.3 \\ &= 0.5\end{aligned} \tag{12.18}$$

因而,2000:Ⅰ通货膨胀的预测值为3.2%+0.5%=3.7%。预测误差较小,为0.4%[①]。看来,增加失业率的多阶滞后项改善了通货膨胀的预测值,它比AR(4)的通货膨胀预测值要好一些。

自回归分布滞后模型。公式(12.16)和公式(12.17)中的模型就是自回归分布滞后(autoregressive distributed lag,简写为ADL)模型:之所以被称为"自回归的",是因为因变量的滞后值作为回归因子被包含进来了;之所以被称为"分布滞后",是因为回归方程中还包含了一个额外预测因子的多阶滞后项("分布滞后")。一般地说,将含有因变量Y_t的p阶滞后和一个额外预测因子X_t的q阶滞后的自回归分布滞后模型,称为ADL(p,q)。用这种符号表示,公式(12.16)中的模型应该是ADL(4,1)模型,而公式(12.17)中的模型应该是ADL(4,4)模型。

自回归分布滞后模型在重要概念12.4中总结。如果有所有这些回归因子,那么公式(12.19)中的符号就显得有点冗长啰嗦,附录12.3根据所谓的滞后算子给出了另一种可选择的符号表示方法。

给定Y和X的所有过去值,ADL模型中的误差项具有条件零均值的假设,即$E(u_t|Y_{t-1},Y_{t-2},\cdots,X_{t-1},X_{t-2},\cdots)=0$,这个假设意味着,在ADL模型中没有包含$Y$或$X$的额外滞后项。换句话说,滞后长度$p$和$q$是真实的滞后长度,其他滞后项的系数为0。



ADL模型包含有因变量的滞后项(自回归成分)和单个额外预测因子X的一个分布滞后项。然而一般地说,可以通过使用多个预测因子来改善预测值。但是,在转到含有多个预测因子的一般时间序列回归模型之前,我们首先介绍平稳性这个概念,它将在后面的讨论中用到。

重要概念12.4

自回归分布滞后模型

含有Y_t的p阶滞后项和X_t的q阶滞后项的自回归分布滞后模型ADL(p,q)是:

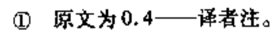
$$Y_t=\beta_0+\beta_1Y_{t-1}+\beta_2Y_{t-2}+\cdots+\beta_pY_{t-p}+\delta_1X_{t-1}+\delta_2X_{t-2}+\cdots+\delta_qX_{t-q}+u_t \tag{12.19}$$

其中,$\beta_0,\beta_1,\cdots,\beta_p,\delta_1,\delta_2,\cdots,\delta_q$是未知系数,$u_t$是满足$E(u_t|Y_{t-1},Y_{t-2},\cdots,X_{t-1},X_{t-2},\cdots)=0$的误差项。

12.4.2 平稳性

时间序列数据的回归分析必然要用来自过去的数据来量化历史关系。如果未来与过去相似,那么就可以用这些历史关系预测未来。但如果未来根本不同于过去,那么这些历史关系可能不是认识未来的可靠指导。

在时间序列回归的范畴中,将历史关系推广到未来这个思想,已经用平稳性(stationarity)概念进行公式化。平稳性的精确定义(在重要概念12.5中给出)是,时间序列



① 原文为0.4——译者注。

变量的分布不随时间的变化而变化。

12.4.3 含有多个预测因子的时间序列回归

含有多个预测因子的一般的时间序列回归模型，是将 ADL 模型加以推广，包含了多个预测因子，以及这些预测因子的滞后项。关于这个模型，在重要概念 12.6 中做了总结。由于多个预测因子及其滞后项的存在，我们必须用双下角标表示回归系数和回归因子。

重要概念 12.5

平稳性

如果一个时间序列 Y_t 的概率分布不随时间的变化而变化，也即如果 $(Y_{s+1}, Y_{s+2}, \cdots, Y_{s+T})$ 的联合分布不依赖于 s，那么它就是平稳的(stationary)，否则，就称 Y_t 是不平稳的(nonstationary)。对于一对时间序列 X_t 和 Y_t，如果 $(X_{s+1}, Y_{s+1}, X_{s+2}, Y_{s+2}, \cdots, X_{s+T}, Y_{s+T})$ 的联合分布不依赖于 s，那么就称他们是联合平稳的(jointly stationary)。平稳性要求，未来与过去至少在概率意义上是相似的。

时间序列回归模型假设。对时间序列数据而言，重要概念 12.6 中的假设修改了截面数据多元回归模型的四个最小二乘假设(见重要概念 5.4)。

第一个假设是，给定所有的回归因子以及包含在回归模型中的滞后项以外的这些回归因子的滞后项，u_t 有条件零均值。这个假设扩展了在 AR 和 ADL 模型中所用到的假设，并隐含着这样的结论，利用 Y 和 X 的所有过去值，Y_t 的最佳预测值由公式(12.20)中的回归给出。

截面数据的第二个最小二乘假设(见重要概念 5.4)是，$(X_{1i}, \cdots, X_{ki}, Y_i)$，$i = 1, \cdots, n$ 是独立同分布的(i.i.d.)。时间序列回归的第二个假设用一个含有两个部分的更恰当的假设代替了 i.i.d. 假设。部分(a)是，数据取自于一个平稳的分布，这样该数据在今天的分布与其在过去的分布相同。这个假设是 i.i.d. 假设的"同分布"部分的时间序列形式：每个抽样为同分布的截面性要求条件被这些变量(包括滞后项)的联合分布不随时间变化而变化的时间序列要求条件所代替。实际上，许多经济时间序列看上去是不平稳的，这意味着这个假设在应用中并不成立。如果时间序列变量是不平稳的，那么在时间序列回归中会出现一个或更多个问题：预测结果可能是有偏的，预测结果可能是无效的(根据同样的数据，可能存在具有较低方差的另外的预测值)，或常规的基于 OLS 的统计推断(例如，通过比较 OLSt 统计量和 ± 1.96 来进行一个假设检验)可能是误导性的。这些问题中具体哪一个会出现，以及它的补救措施是什么，都依赖于该非平稳性的来源。在 12.6 节和 12.7 节中，我们研究了在经济时间序列实证研究中存在的两类重要的非平稳性——趋势和突变——产生的原因和解决的方法。不过，目前我们还是简单地假设所研究的序列是联合平稳的，从而我们集中研究平稳性变量的回归问题。

重要概念 12.6

含有多个预测因子的时间序列回归



一般的时间序列回归模型允许有 k 个额外的预测因子，其中包括第一个预测因子的 q_1 阶滞后，第二个预测因子的 q_2 阶滞后，依此类推。

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \beta_2 Y_{t-2} + \cdots + \beta_p Y_{t-p} + \delta_{11} X_{1t-1} + \delta_{12} X_{1t-2} + \cdots + \delta_{1q_1} X_{1t-q_1} + \cdots +$$

$$\delta_{k1} X_{kt-1} + \delta_{k2} X_{kt-2} + \cdots + \delta_{kq_k} Z_{kt-q_k} + u_t \quad (12.20)$$

其中：

1. $E(u_t|Y_{t-1},Y_{t-2},\cdots,X_{1t-1},X_{1t-2},\cdots,X_{kt-1},X_{kt-2},\cdots)=0$。

2. (a)随机变量$(Y_t,X_{1t},\cdots,X_{kt})$为平稳分布；(b)随着 j 值的增大，$(Y_t,X_{1t},\cdots,X_{kt})$与$(Y_{t-j},X_{1t-j},\cdots,X_{kt-j})$变成独立的。

3. $X_{1t},\cdots,X_{kt}$和 Y_t 具有非零的有限的四阶矩。

4. 不存在完全多重共线性。

第二个假设的部分(b)要求，当分割时间的时期数变得很大时，所研究的随机变量变成独立分布的。这个假设用时间序列的要求条件——当观测值被长时期分隔时它们是独立分布的——代替了变量在不同的观测值之间是独立分布的这个截面性数据要求的条件。该假设有时被称为**弱依存性**(weak dependence)，它确保在大样本条件下，数据中存在充分的随机性使得大数定律和中心极限定理成立。这里我们不给出弱依赖性条件精确的数学表述，感兴趣的读者可参考 Hayashi(2000，第 2 章)。

第三个假设(与截面数据的第三个最小二乘假设相同)是，所有变量具有非零的有限的四阶矩。

最后，第四个假设(它也和截面数据的假设相同)是，回归因子不是完全多重共线的。

统计推断和格兰杰因果关系检验。在重要概念 12.6 的假设下，对回归系数使用 OLS 所做的统计推断，按照使用截面数据通常所采用的相同方法来执行。

F 统计量在时间序列预测中的一个很有用的应用是，检验某一个在方程中所包含的回归因子的滞后项是否比模型中其他的回归因子更有预测意义。变量没有预测意义的主张，与这个变量所有的滞后项的系数都为零的零假设是对应的。检验这个零假设的 F 统计量被称为**格兰杰因果关系统计量**(Granger causality statistic)，相关的检验被称为**格兰杰因果关系检验**(Granger causality test)(Granger(1969))。这个检验的内容在重要概念 12.7 中总结。

重要概念 12.7

格兰杰因果关系检验(是否有预测意义的检验)

格兰杰因果关系统计量，就是检验公式(12.20)中某一变量的所有系数值(例如，X_{1t-1}，X_{1t-2}，$\cdots$，X_{1t-q_1}的系数)都为 0 的这个零假设的 F 统计量。这个零假设意味着，这些回归因子对 Y_t 没有预测意义(包含在其他回归因子中的预测内容除外)，对这个零假设的检验就被称为格兰杰因果关系检验。

格兰杰因果关系与本书中其他部分所提到的因果关系几乎没有什么联系。在第 1 章，因果关系是根据一个理想的随机化控制实验来定义的，其中在实验中使用 X 的不同值，然后观察它们对 Y 的随后的影响。相反，格兰杰因果关系意味着，给定回归模型中的其他变量，如果 X 格兰杰引致 Y，那么 X 就是 Y 的一个有用的预测因子。尽管"格兰杰预测能力"是一个比"格兰杰因果关系"更准确的术语，但后者已成为经济计量学专业术语的一部分。

举例来说，考虑通货膨胀率的变化与其过去值和失业率的过去值之间的关系。根据公式(12.17)中的 OLS 估计值，检验失业率的四个滞后项系数都为 0 的这个零假设的 F 统计量为 8.51($p<0.001$)。按照重要概念 12.7 中的专业术语，我们能够得出结论(在 1% 的显著性水平下)，即该失业率格兰杰引致通货膨胀率的变化，但这并不一定意味着失业率的变化会导致——在第 1 章的意义下——通货膨胀率的随后变化。可是这确实意味着，除了通

货膨胀率的过去值所包含的有用信息外，失业率的过去值看起来包含了对预测通货膨胀率变化有用的信息。

12.4.4 预测的不确定性和预测的区间

在任何估计问题中，报告估计值不确定性的一个测量指标是一个很好的习惯，预测也不例外。测量一项预测不确定性的一个指标是均方根预测误差。在误差项 u_t 服从正态分布的这个额外假设下，可用 RSMFE 构造一个预测区间，即以一定的概率包含该变量未来值的一个区间。

预测的不确定性。预测误差由两部分组成：由回归系数的估计所引起的不确定性，和与 u_t 的将来未知值有关的不确定性。对于系数很少但观测值很多的回归而言，由未来的 u_t 所引起的不确定性可能远大于和参数估计有关的不确定性。不过，一般地讲，这两个不确定性的来源都是重要的，因此我们现在为 RMSFE 设计一个把这两个不确定性的来源都考虑进来的表达式。

为了使符号简单，考虑含有单个预测因子的一个 ADL(1,1)模型的 Y_{T+1} 的预测问题，即 $Y_t=\beta_0+\beta_1Y_{t-1}+\delta_1X_{t-1}+u_t$，并假设 u_t 是同方差的，预测值为 $\hat{Y}_{T+1|T}=\hat{\beta}_0+\hat{\beta}_1Y_T+\hat{\delta}_1X_T$，预测误差为：

$$Y_{T+1}-\hat{Y}_{T+1|T}=u_{T+1}-[(\hat{\beta}_0-\beta_0)+(\hat{\beta}_1-\beta_1)Y_T+(\hat{\delta}_1-\delta_1)X_T] \tag{12.21}$$

由于 u_{T+1} 具有条件零均值且是同方差的，因此 u_{T+1} 的方差为 σ_u^2，它与公式(12.21) 中最后那个带括号的表达式不相关。所以，均方预测误差(MSFE)为：



$$\begin{aligned}\text{MSFE}&=E[(Y_{T+1}-\hat{Y}_{T+1|T})^2]\\&=\sigma_u^2+\text{var}[(\hat{\beta}_0-\beta_0)+(\hat{\beta}_1-\beta_1)Y_T+(\hat{\delta}_1-\delta_1)X_T]\end{aligned} \tag{12.22}$$

而 RMSFE 是 MSFE 的平方根。

对 MSFE 的估计需要估计公式(12.22)中的两个部分。第一项 σ_u^2 可以用该回归的标准误的平方来估计，见 12.3 节中的讨论内容。第二项需要估计该回归系数的一个加权平均的方差，估计的方法在 6.1 节中已论述过了(见公式(6.7)之后的论述)。

估计 MSFE 的另一种可供选择的方法是使用伪样本外预测的方差(the variance of pseudo out-of-sample forecasts)，这是 12.7 节中所讨论的一种方法。

预测的区间。一个预测区间很像一个置信区间，只是它是对一项预测而言的。也就是说，一个 95% 的预测区间(forecast interval)是指在 95% 的重复应用中包含该序列未来值的一个区间。

预测区间和置信区间之间的一个重要差异是，通常一个 95% 的置信区间的计算公式(估计量 ±1.96 倍标准误)已由中心极限定理所证明，因此它对误差项的一个更大范围的分布都成立。相反，由于公式(12.21)中的预测误差包含误差项 u_{T+1} 的未来值，因此为了计算预测区间，需要估计该误差项的分布或对那个分布作一些假设。

在实际中，假设 u_{T+1} 服从正态分布是很方便的。如果是这样的话，适用于 $\hat{\beta}_0$、$\hat{\beta}_1$ 和 $\hat{\delta}_1$ 的公式(12.21)和中心极限定理就意味着，该预测误差是两个独立的正态分布项的和，因此，该预测误差本身也服从方差为 MSFE 的正态分布。由此可得，一个 95% 的置信区间由 $\hat{Y}_{T+1|T}\pm1.96SE(Y_{T+1}-\hat{Y}_{T+1|T})$ 给出，其中，$SE(Y_{T+1}-\hat{Y}_{T+1|T})$ 是 RMSFE 的一个估计量。



这个讨论集中分析了误差项 u_{T+1} 是同方差的情形。如果 u_{T+1} 是异方差的，那么就需要开发出异方差的模型，使得给定 Y 和 X 最近期的值，能够估计出公式(12.22)中的 σ_u^2 项。

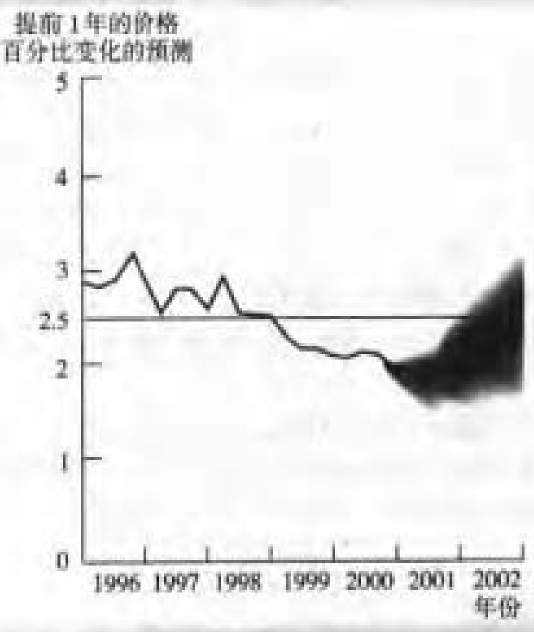
对这种条件异方差的建模方法在 14.5 节中讨论。

由于未来事件的不确定性，即关于 u_{T+1} 的不确定性，95% 的预测区间可能太宽，这样它们在决策中的用途就很有限，因此，专业预测人士经常报告比 95% 更窄的预测区间，比如说一个标准误的预测区间（如果误差项服从正态分布，那么它就是 68% 的预测区间）。另一种选择是，一些预测人士报告多个预测区间，就像英格兰银行的经济学家们在公布通货膨胀预测值时所做的那样（见本章的一般兴趣框）。

一般兴趣框

血　河

作为货币政策决策中向公众报告的一部分，英格兰银行定期公布通货膨胀的预测值。这些预测结果，将该银行中专业经济计量学家所主张的经济计量模型中所得到的结果，与该银行高级机构的成员和货币政策委员会成员的专家判断结合在一起。该预测结果以一系列的预测区间的形式给出，这些预测区间被设计成反映这些经济学家认为通货膨胀可能走向的大概范围。在他们的通货膨胀报告中，银行用红色子体标示这些范围，其中深红色留给中心的范围线。尽管该银行平淡地称之为"扇形图"，但新闻界却将这些扩散的红色阴影叫做"血河"。



注：2001 年 2 月份，英格兰银行的扇形图显示了通货膨胀的预测区间。

图 12—4　血河

图 12—4 显示了 2001 年 2 月份的血河（在这个图形中血是绿色的，不是红色的，因此需要你发挥一下想象力）。这个图形，也即 2001 年 2 月份的图形说明，该银行的经济学家们预测下一年的通货膨胀率将从大约 3% 下降到略超过 2%，随后又会增加。但这个预测存在相当大的不确定性。在他们的书面论述中，他们特别引用了美国通货膨胀进一步下降的可能性——事实上，美国经济在 2001 年是衰退的——它会导致英国通货膨胀率的下降。英国的通货膨胀率确实下降了，这说明他们的预测是一个很好的预测：在 2001 年的第四季度，通货膨胀率是 2.0%。

在向更公开化的中央银行迈进的运动中，英格兰银行是个先驱，其他中央银行现在也公

布了通货膨胀预测值。由货币政策制定者所做的决策是个困难的决策,并且影响到许多社会大众的生活和钱包。在一个信息时代的民主政治下,英格兰银行的经济学家们更加理性化,对公众而言,理解银行的经济展望报告和困难决策背后的推理是特别重要的。

要看到用最初的红色显示的血河,请访问英格兰银行的网站 www.bankofengland.co.uk/inflationreport。

12.5 利用信息准则选择滞后长度

在12.3节和12.4节中所估计的通货膨胀回归含有预测因子的一阶或四阶的滞后项。一阶滞后项有一定的意义,但为什么四阶滞后项也有意义?更一般地讲,在时间序列回归中应包含多少阶滞后期?本节就讨论选择滞后阶数的统计方法,首先我们在自回归中讨论,然后在含有多个预测因子的时间序列回归模型中讨论。

12.5.1 确定自回归的阶数

在实际中,选择一个自回归的阶数 p 需要在包括更多滞后项所带来的好处与额外的估计不确定性的成本之间进行权衡。一方面,如果一个估计的自回归的阶数太低,那么你就会遗漏包含在更远期滞后值中的潜在的有价值的信息。另一方面,如果阶数太高,你又会估计比必要的数目更多的系数,这又会将额外的估计误差引入到你的预测值中。

F 统计量方法。选择 p 的一种方法是,从一个含有多阶滞后项的模型开始,并对最后那个滞后项进行假设检验。例如,你可以从估计 AR(6)模型开始,并检验第六阶滞后项的系数在5%的水平下是否显著;如果不显著,就剔除它,并估计 AR(5)模型,检验第五阶滞后项系数,依此类推。这个方法的缺点是它会生成一个过大的模型,至少有时候会:如果真正的 AR 阶数是5,那么第六阶的系数就是0,使用 t 统计量的一个5%水平的检验只有5%的偶然机会会不正确地拒绝这个零假设。因此,当 p 的真实值为5时,这种方法会有5%的机率将 p 估计为6。

BIC 准则。围绕这个问题的一种方法是用最小化"信息准则"来估计 p。这种信息准则之一就是贝叶斯信息准则(Bayes information criterion,简写为 BIC),也被称为许瓦兹信息准则(Schwarz information criterion,简写为 SIC),它的公式是:

$$\mathrm{BIC}(p)=\ln\left[\frac{SSR(p)}{T}\right]+(p+1)\frac{\ln T}{T} \tag{12.23}$$

其中,$SSR(p)$ 是所估计的 AR(p)的残差平方和。p 的 BIC 估计量 $\hat{p}$ 是在可能选择的 $p=0,1,\cdots,p_{\max}$ 范围内使 BIC(p)最小的那个值,其中 $p_{\max}$ 是所考虑的 p 中最大的那个值。

BIC 公式乍看起来可能有一点神秘,但它具有直观的吸引力。考虑公式(12.23)中的第一项,由于回归系数是用 OLS 估计的,因此当增加滞后项时,残差平方和必然会减少(或至少不会增加)。相反,第二项是所估计的回归系数的个数(滞后阶数 p 加上截距项1)乘上因子$(\ln T)/T$。当增加滞后项时,第二项会增大。BIC 要平衡这两个影响力量,使得最小化 BIC 的滞后阶数是真实滞后长度的一个一致估计量。这个论述的数学证明在附录12.5中给出。

作为一个例子,考虑估计通货膨胀率变化的 AR 的阶数。对于最大阶数为6($p_{\max}=6$)的自回归而言,BIC 的不同计算步骤在表12—3中完成。例如,对公式(12.7)中的 AR(1)模型而言,$SSR(1)/T=2.726$,因此 $\ln[SSR(1)/T]=1.003$。由于 $T=152$(38年,每年4个

计额外系数的成本之间进行权衡。

F 统计量方法。和在单变量自回归中一样，确定要包括的滞后阶数的一种方法就是用 *F* 统计量检验系数集都等于零的联合假设。例如，在公式(12.17)的讨论中，我们检验了失业率的第二阶到第四阶滞后项的系数都等于零的这个零假设，备择假设是它们都不为零。这个假设在1%的显著性水平下被拒绝，这为较长的滞后设定提供了支持。如果所比较的模型数量较少，那么这个 *F* 统计量方法易于使用。然而一般来说，*F* 统计量方法会生成过大的模型，从这个意义上说，真实的滞后阶数被高估了。

信息准则。和自回归一样，BIC 和 AIC 也可用于估计含有多个预测因子的时间序列模型中滞后项和变量的个数。如果回归模型中有 *K* 个系数(包括截距)，那么 BIC 为：

$$\mathrm{BIC}(K)=\ln\left[\frac{SSR(K)}{T}\right]+K\frac{\ln T}{T} \tag{12.25}$$

AIC 可以同样的方式定义，但用 2 代替公式(12.25)中的 $\ln(T)$。对每个所要评价的模型，可以求出 BIC(或 AIC)的值，根据信息准则，具有 BIC(或 AIC)最小值的模型便是我们所要选择的模型。

使用信息准则估计滞后长度时，有两个重要的实际情况要考虑。第一，与自回归的情况一样，所有候选模型必须用相同的样本进行估计；用公式(12.25)的符号，用于估计模型的观测期数 *T* 对所有模型必须相同。第二，当存在多个预测因子时，由于需要计算许多不同的模型(许多滞后参数的组合)，因此在计算上这种方法的工作量很大。在实践中，一个捷径就是要求所有回归因子具有相同的滞后阶数，即要求 $p=q_1=\cdots=q_k$，因此，只有 $p_{\max}+1$ 个模型需要比较(对应于 $p=0,1,\cdots,p_{\max}$)。

第4部分

12.6 非平稳性 I：趋势

在重要概念 12.6 中，假设因变量和回归因子都是平稳的。如果事实并非如此，也就是说，如果因变量和/或回归因子是不平稳的，那么常规的假设检验、置信区间和预测则可能是不可靠的。由非平稳性所引起的准确性问题及其解决这个问题的方法，依赖于该非平稳性的性质。

在这一节和下一节里，我们研究经济时间序列数据中两类重要的非平稳性：趋势和突变。在每一节里，我们首先介绍非平稳性的性质，然后讨论如果这类非平稳性存在却被忽略的情况下对时间序列回归所产生的后果。接下来我们提出非平稳性的检验，讨论由那个特殊类型的非平稳性所引起的问题的补救措施或解决方法。我们从讨论趋势开始。

12.6.1 什么是趋势？

趋势(trend)是指某一变量随时间变化所产生的持续的长期的运动。某一时间序列变量围绕着它的趋势波动。

检查图 12—1(a)我们发现，美国通货膨胀率包含从开始到 1982 年的一个总体上升趋势和之后的一个总体下降趋势。图 12—2(a)、12—2(b)和 12—2(c)中的序列也有趋势，但它们的趋势大不相同。美国联邦基金利率的趋势类似于美国通货膨胀率的趋势。在 1972 年固定汇率体系崩溃以后，美元/英镑汇率明显具有一个延长的下降趋势。日本实际 GDP 的对数的趋势较为复杂：起初快速增长，而后适度增长，最后缓慢增长。

确定性趋势和随机性趋势。在时间序列数据中存在两类可见到的趋势,即确定性趋势和随机性趋势。确定性趋势(deterministic trend)是时间的一个非随机函数。例如,一个确定性趋势可能是时间的线性函数。如果通货膨胀拥有一个确定性的线性趋势,使得它每个季度增加0.1个百分点,那么这个趋势可以记作$0.1t$,其中t为季度测量。相反,一个随机性趋势(stochastic trend)是随机的,而且随时间的变化而变化。例如,通货膨胀中的一个随机趋势可能先表现出一个延长期的上升,随后表现出一个延长期的下降,就像图12—1中的通货膨胀趋势一样。

和许多经济计量学家一样,我们认为将经济时间序列建立为含有随机性趋势而不是确定性趋势的模型更合适。经济学是个很复杂的事物,很难使一个确定性趋势所隐含的预测结果与工人、商业和政府所面对的年复一年的复杂因素与意外事件相一致。例如,尽管美国通货膨胀一直上涨到20世纪70年代,但这既不会注定永远地上涨,也不会注定又一次地下降。然而,通货膨胀的缓慢上涨现在可被理解为由于坏运气和不良的货币政策而发生的,缓慢上涨大部分归因于联邦储备管理委员会的强硬决策的结果。同样,从1972年到1985年,美元/英镑汇率趋势向下,随后向上漂移,但这些运动也是复杂经济力量作用的结果。由于这些力量的变化不可预测,因此这些趋势被看做是含有大量的不可预测或随机的成分,这是很有意义的。

由于这些原因,我们对经济时间序列中的趋势的处理集中在随机性趋势上,而不是集中在确定性趋势上。在时间序列数据中,当我们提到"趋势"时,我们是指随机性趋势,除非我们特别说明。这一节给出最简单的随机趋势模型,即随机游动模型,其他的趋势模型在14.3节中讨论。



趋势的随机游动模型。含有随机趋势的一个变量最简单的模型是随机游动模型。如果某一时间序列Y_t的变化是独立同分布的,也就是说,如果:

$$Y_t = Y_{t-1} + u_t \tag{12.26}$$

那么就称Y_t服从随机游动(random walk),其中,u_t是独立同分布的。不过,一般情况下,我们将使用术语"随机游动"来指那些满足公式(12.26)的时间序列,这里u_t具有条件零均值,即$E(u_t|Y_{t-1},Y_{t-2},\cdots)=0$。

随机游动的基本思想是:一个序列的明天值,等于它的今天值加上一个不可预测的变化。由于Y_t所遵循的路径由随机的"步"u_t组成,因此那条路径就是"随机游动"。根据从开始到$t-1$时刻的数据,Y_t的条件均值是Y_{t-1},也就是说,因为$E(u_t|Y_{t-1},Y_{t-2},\cdots)=0$,所以$E(Y_t|Y_{t-1},Y_{t-2},\cdots)=Y_{t-1}$。换句话说,如果$Y_t$服从随机游动,那么明天值的最优预测是它的今天值。

一些序列,如图12—2(c)中日本GDP的对数,具有明显的上升趋势,在这种情况下,对该序列的最优预测必须包含对该序列上升趋势的一个调整。这种调整使随机游动模型扩展为包括向一个方向运动或者向另一个方向运动的趋势运动项或"漂移项",这个扩展形式被称为带漂移项的随机游动(random walk with drift)。

$$Y_t = \beta_0 + Y_{t-1} + u_t \tag{12.27}$$

其中,$E(u_t|Y_{t-1},Y_{t-2},\cdots)=0$,$\beta_0$为随机游动的"漂移项"。如果$\beta_0$是正的,那么平均来说$Y_t$会增加。在带漂移项的随机游动模型中,对序列明天值的最优预测,是今天的序列值加上漂移项β_0。

随机游动模型(适当时带漂移项)至今仍是简单通用的,并且是本书中所使用的测定趋势的主要模型。

随机游动是非平稳的。如果 Y_t 服从随机游动，那么它就不是平稳的。一项随机游动的方差随时间而增加，因此，Y_t 的分布随时间而变化。理解这一点的一种方法就是认识到，在公式(12.26)中，由于 u_t 序列无关，因此 $\mathrm{var}(Y_t)=\mathrm{var}(Y_{t-1})+\mathrm{var}(u_t)$；由于 Y_t 是平稳的，$\mathrm{var}(Y_t)$ 不可能依赖于时间，因此，$\mathrm{var}(Y_t)=\mathrm{var}(Y_{t-1})$ 必然成立，但这只有在 $\mathrm{var}(u_t)=0$ 时才会发生。理解这一点的另一种方法是假想 Y_t 从 0 开始，即 $Y_0=0$，那么 $Y_1=u_1$，$Y_2=u_1+u_2$，依此类推，这样 $Y_t=u_1+u_2+\cdots+u_t$。由于 u_t 序列无关，因此，$\mathrm{var}(Y_t)=\mathrm{var}(u_1+u_2+\cdots+u_t)=t\sigma_u^2$。因而，$Y_t$ 的方差依赖于 t，事实上，它随着 t 的增加而增加。由于 Y_t 的方差依赖于 t，因此它的分布依赖于 t，即它是非平稳的。

因为随机游动的方差无限地增加，所以它的总体自相关没有定义（一阶自协方差和方差是无限的，二者的比值没有意义）。不过，随机游动的一个特征就是它的样本自相关系数倾向于非常逼近于 1，事实上，随机游动的 j 阶样本自相关系数依概率收敛于 1。

随机趋势、自回归模型和单位根。随机游动模型是 AR(1)模型（公式(12.8)）的一个特例，其中 $\beta_1=1$。换句话说，如果 Y_t 满足 $\beta_1=1$ 的 AR(1)，那么 Y_t 包含一个随机趋势，并且是非平稳的。然而，如果 $|\beta_1|<1$，且 u_t 是平稳的，那么 Y_t 和它的滞后项的联合分布不依赖于 t（结果在附录 12.2 中给出）。因此只要 u_t 是平稳的，Y_t 就是平稳的。

AR(p)平稳的类似条件比 AR(1)的条件 $|\beta_1|<1$ 更复杂。它的正式陈述涉及多项式 $1-\beta_1z-\beta_2z^2-\beta_3z^3-\cdots-\beta_pz^p$ 的根（这个多项式的根就是方程 $1-\beta_1z-\beta_2z^2-\beta_3z^3-\cdots-\beta_pz^p=0$ 的解）。对于一个平稳的 AR(p)，这个多项式的根的绝对值必定都大于 1。在 AR(1)的特殊情况下，根是解 $1-\beta_1z=0$ 所得到的 z 的值，故它的根是 $z=1/\beta_1$。因而，根的绝对值大于 1 的命题等价于 $|\beta_1|<1$。

如果 AR(p)有个等于 1 的根，则称该序列有单位自回归根（unit autoregressive root），或更简单地称为单位根（unit root）。如果 Y_t 有单位根，那么它就包含随机性趋势。如果 Y_t 是平稳的（因而没有单位根），它就不包含随机性趋势。由于这个原因，我们会交替使用术语“随机趋势”和“单位根”。

12.6.2 随机性趋势所引致的问题

如果一个回归因子含有随机性趋势（含有单位根），那么即使在大样本条件下，它的系数的 OLS 估计量及其 OLSt 统计量可能会有非标准的（即非正态的）分布。我们讨论这个问题的三个具体的方面：第一，如果自回归系数的真实值是 1，那么 AR(1)中它的估计量偏向于 0；第二，即使在大样本条件下，具有随机性趋势的回归因子的 t 统计量可能服从非正态分布；第三，由随机性趋势所引致的风险中，一个极端例子就是，两个独立的序列如果都有随机性趋势的话，它们会以很高的概率使人误认为是相关的，这种情况被称作伪回归。

问题 1：偏向于 0 的自回归系数。假设 Y_t 服从公式(12.26)中的随机游动，但这对经济计量学家是未知的，他们反而估计公式(12.8)中的 AR(1)模型。由于 Y_t 是非平稳的，重要概念 12.6 中时间序列回归的最小二乘假设不成立，因而，作为一个一般的情况，我们不能依靠通常服从大样本正态分布的估计量和检验统计量。事实上，在这个例子中，自回归系数的 OLS 估计量 $\hat{\beta}_1$ 是一致的，但它服从非正态分布，即使在大样本条件下也是如此，$\hat{\beta}_1$ 的渐近分布偏向于 0。$\hat{\beta}_1$ 的期望值约为 $E(\hat{\beta}_1)=1-5.3/T$。对一般经济应用中的样本容量来说，这会导致大的偏差。例如，20 年的季度数据包含 80 个观察值，在此情况下，$\hat{\beta}_1$ 的期望值是 $E(\hat{\beta}_1)=1-5.3/80=0.934$。此外，这个分布有很长的左尾。$\hat{\beta}_1$ 的 5% 百分位数约为 1 -

14.1/T，对 $T=80$ 而言，相应地得到0.824，因此有5%的机会 $\hat{\beta}_1<0.824$。

这个偏向于0的偏差的一个含义是：如果 Y_t 服从随机游动，那么基于AR(1)模型所得出的预测，可能比基于将真实值设定为 $\beta_1=1$ 的随机游动模型所做出的预测，结果会更糟糕。这个结论也适用于更高阶的自回归情形，在那里当序列确实含有单位根时，强加一个单位根（也就是说，用一阶差分代替水平值估计自回归）对预测有好处。

问题2：t 统计量的非正态分布。如果一个回归因子有随机趋势，那么通常的OLSt 统计量在零假设下服从非正态分布，即使在大样本情况下也是如此。这个非正态分布意味着，常规的置信区间不是有效的，而且不能像往常那样进行假设检验。一般地说，这个 t 统计量的分布不容易列表显示，因为该分布依赖于我们所研究的那个回归因子和其他回归因子之间的关系。可能列表显示这个分布的一个重要情形是在含有单位根的自回归的情况下，当我们开始检验时间序列是否含有随机趋势时，我们还会回到这个特殊情况中。

问题3：伪回归。随机性趋势会使两个不相关的时间序列看上去相关，这个问题被称为**伪回归**(spurious regression)。

例如，从20世纪60年代中期到80年代初期，美国通货膨胀逐步上升，同时日本GDP也逐步上升。使用常规的测量，这两个趋势凑合起来生成一个看上去是"显著的"回归。使用OLS利用从1965年到1981年的数据进行估计，这个估计的回归是：

$$\widehat{U.S.\ Inflation}_t = \underset{(0.08)}{-2.84} + \underset{(0.02)}{0.18}\ JapaneseGDP_t, \bar{R}^2=0.56 \tag{12.28}$$

斜率系数的 t 统计量超过9，根据我们常用的标准，这表明两个序列之间有强正向关系，且 $\bar{R}^2$ 也较高。但使用从1982年到1999年的数据进行同样的回归，我们得到：

$$\widehat{U.S.\ Inflation}_t = \underset{(1.37)}{6.25} - \underset{(0.01)}{0.03}\ JapaneseGDP_t, \bar{R}^2=0.07 \tag{12.29}$$

公式(12.28)和公式(12.29)中的回归简直太不相同了。按照字面解释，方程(12.28)显示了强正向关系，而公式(12.29)显示了弱负向关系。

这两个矛盾结果的根源就是这两个序列都有随机性趋势。这些趋势恰巧从1965年到1981年是一致的，但从1982年到1999年却不一致了。实际上，在经济方面或政治方面都不存在有说服力的理由认为这两个序列中的趋势是相关的。简而言之，这些回归是虚假（或伪）回归。

公式(12.28)和公式(12.29)中的回归在实证上说明了这样的理论观点：当序列包含随机性趋势时，OLS估计量可能会产生误导（见练习12.6的计算机模拟，它证明了这个结论）。一种特殊情况是当两个序列的趋势成分相同的时候，即当序列包含相同的随机性趋势的时候，某些以回归为基础的方法是可靠的。如果是这样的话，那么就说该序列是协整的。识别和分析协整经济时间序列的方法在14.4节中讨论。

12.6.3 随机趋势的识别：单位AR根检验

时间序列数据中的趋势可以用正式的和非正式的方法来识别。非正式方法包括审视时间序列的数据图和计算自相关系数，就像我们在12.2节中所做的那样。因为如果序列有随机性趋势，那么一阶自相关系数会接近于1，至少在大样本情况下是如此，所以，一个小的一阶自相关系数和一个没有明显趋势的时间序列图结合在一起表明了该序列确实没有趋势。然而，如果还存在疑问，那么还有正式的统计方法用来检验序列中存在随机性趋势这一零假设对应于序列中不存在随机性趋势这一备择假设。

在本节中，我们用迪基—富勒检验（以其创始人 David Dickey 和 Wayne Fuller(1979)的名字命名）检验随机性趋势。尽管迪基—富勒检验不是随机性趋势的惟一检验方法（另一个检验在 14.3 节中讨论），但它是实际中最常用的检验，并且是最可靠的检验之一。

AR(1)模型中的迪基—富勒检验。迪基—富勒检验（Dickey-Fuller test）的出发点是自回归模型。如前面所讨论的，公式(12.27)中的随机游动是 $\beta_1=1$ 的 AR(1)模型的一个特例。如果 $\beta_1=1$，那么 Y_t 是非平稳的且含有一个（随机）趋势。因而，在 AR(1)模型内，Y_t 含有随机性趋势的假设可以通过检验以下的假设进行检验。

$$在\ Y_t=\beta_0+\beta_1 Y_{t-1}+u_t 中, H_0:\beta_1=1, H_1:\beta_1<1 \tag{12.30}$$

如果 $\beta_1=1$，那么 AR(1)有值为 1 的自回归根，因此，公式(12.30)中的零假设是AR(1)有单位根，备择假设是它是平稳的。

通过将公式(12.30)两边都减去 Y_{t-1} 所得到的修正的模型，对这个模型执行检验最容易。令 $\delta=\beta_1-1$，那么公式(12.30)变为：

$$在\ \Delta Y_t=\beta_0+\delta Y_{t-1}+u_t 中, H_0:\delta=0, H_1:\delta<0 \tag{12.31}$$

在公式(12.31)中，检验 $\delta=0$ 的 OLSt 统计量被称为迪基—富勒统计量（Dickey-Fuller statistic）。公式(12.31)中的表达式是简便的，因为回归软件会自动给出检验 $\delta=0$ 的 t 统计量。注意，迪基—富勒检验是单边的，因为相关的备择假设为 Y_t 是平稳的，所以有 $\beta_1<1$、或者 $\delta<0$。用"非稳健的"标准误，即附录 4.4 中所介绍的"仅适用于同方差"的标准误（公式(4.62)中单个回归因子情形和 16.4 节中多元回归模型的情形）计算迪基—富勒统计量[①]。

第
4
部
分

AR(p)模型中的迪基—富勒检验。在公式(12.31)的意义下所给出的迪基—富勒统计量只适用于 AR(1)模型。如 12.3 节中所讨论的，对一些序列而言，AR(1)模型没有捕捉到 Y_t 中所有的序列相关，在此情况下，选择一个较高阶的自回归会更合适。

AR(p)模型的迪基—富勒检验的扩展形式在重要概念 12.8 中总结。在零假设下，$\delta=0$，ΔY_t 是平稳的 AR(p)。在备择假设下，$\delta<0$，因此 Y_t 是平稳的。因为用于计算该迪基—富勒统计量的这种形式的回归被 ΔY_t 的滞后项扩大了，所以相应的 t 统计量又被称为增项的迪基—富勒（ADF）统计量（augmented Dickey-Fuller (ADF) statistic）。

一般地说，滞后长度 p 是未知的，但可以使用应用于不同 p 值的公式(12.32)的回归的信息准则来估计它。对 ADF 统计量的研究表明，含有太多的滞后项比含有太少的滞后项要好，因此对 ADF 统计量而言，建议用 AIC 代替 BIC 来估计 p。[②]

重要概念 12.8

单位自回归根的增项的迪基—富勒检验

单位自回归根的增项的迪基—富勒检验，是检验回归方程(12.32)中零假设 $H_0:\delta=0$ 和对应的单边备择假设 $H_1:\delta<0$。

$$\Delta Y_t=\beta_0+\delta Y_{t-1}+\gamma_1\Delta Y_{t-1}+\gamma_2\Delta Y_{t-2}+\cdots+\gamma_p\Delta Y_{t-p}+u_t \tag{12.32}$$

在该零假设下，Y_t 有随机性趋势；在备择假设下，Y_t 是平稳的。这里的 ADF 统计量就是检验公式(12.32)中 $\delta=0$ 的 OLSt 统计量。

相反，如果备择假设是"Y_t 围绕确定性的线性时间趋势是平稳的"，那么，这个趋势"t"



① 在单位根零假设下，通常的"非稳健"标准误会生成一个实际上对异方差稳健的 t 统计量，一个令人惊讶而又特殊的结论。

② 请参考 Stock(1994)关于迪基—富勒和其他单位根检验统计量有限样本性质的相关模拟研究综述。

$$\widehat{\Delta Inf_t} = \underset{(0.23)}{0.53} - \underset{(0.04)}{0.11}\, Inf_{t-1} - \underset{(0.08)}{0.14}\, \Delta Inf_{t-1} - \underset{(0.08)}{0.25}\, \Delta Inf_{t-2} + \underset{(0.08)}{0.24}\, \Delta Inf_{t-3} + \underset{(0.08)}{0.01}\, \Delta Inf_{t-4} \quad (12.34)$$

这里的 ADFt 统计量就是检验假设 Inf_{t-1} 的系数为 0 的 t 统计量，这里 $t = -2.60$。根据表 12—4，5% 的临界值为 -2.86。因为 ADFt 统计量 -2.60 比 -2.86 大，所以该检验在 5% 的显著性水平下不能被拒绝。因此，根据公式(12.34)中的回归，我们(在 5% 的水平下)在通货膨胀是平稳的备择假设下不能拒绝通货膨胀有单位根的零假设，也就是说，通货膨胀包含随机性趋势，对应的备择假设是该序列是平稳的。

公式(12.34)中的 ADF 回归包括了 ΔInf_t 的四阶滞后项来计算 ADF 统计量。当用 AIC 估计滞后阶数时，其中 $0 \leqslant p \leqslant 6$，滞后长度的 AIC 估计量却是 3。当使用三阶滞后时(即当 ΔInf_{t-1}，ΔInf_{t-2} 和 ΔInf_{t-3} 作为回归因子被包括进来时)，ADF 统计量为 -2.65，它比 -2.86 大，因而，当 ADF 回归中的滞后阶数用 AIC 选择时，通货膨胀含有随机性趋势的假设在 5% 的显著性水平下不能被拒绝。

这些检验是在 5% 的显著性水平下执行的。不过，在 10% 的显著性水平下该检验也拒绝了单位根的零假设：-2.60(四阶滞后)与 -2.65(三阶滞后)的 ADF 统计量比 -2.57 的 10% 临界值稍微小一点。因此，ADF 统计量做了一个意思相当含糊的描述，预测者必须以一个可靠的信息为基础做出判断，是否将通货膨胀建立为一个含有随机性趋势的模型。显然，图 12—1(a)中的通货膨胀表现出长期摆动，与随机性趋势模型一致。此外，在实际中，许多预测者把美国通货膨胀看做为含有随机性趋势，这里我们遵循这种策略。

12.6.4 避免由随机性趋势所引致的问题

处理一个序列中的趋势的最可靠的方法，就是对该序列进行变换，使得它不含有趋势。如果序列含有趋势，即序列有单位根，那么该序列的一阶差分就没有趋势。例如，如果 Y_t 服从随机游动，则有 $Y_t = \beta_0 + Y_{t-1} + u_t$，那么 $\Delta Y_t = \beta_0 + u_t$ 是平稳的。因此，使用一阶差分就剔除了序列中的随机游动趋势。

实际上，你很少能够确定一个序列中是否含有随机性趋势。作为一个一般的观点，回想一下，无法拒绝零假设并不一定意味着这个零假设是真的，确切地说，它仅仅意味着你没有充分的证据得出它是假的结论。因此，用 ADF 统计量无法拒绝存在单位根的零假设并不意味着该序列确实有单位根。例如，在一个 AR(1)模型中，真实的系数 β_1 可能非常逼近于 1，比如说 0.98，在此情况下 ADF 检验具有很低的效力，也就是说，就我们的通货膨胀序列的样本规模，正确地拒绝该零假设的概率很低。即使无法拒绝存在单位根的零假设，也并不意味着序列有单位根，真实的自回归根近似等于 1 仍然是合理的，因此应该使用序列的差分而非水平值①。

12.7 非平稳性Ⅱ：突变

当总体回归函数在样本期间发生变化时，便出现了第二类非平稳性。在经济学中，发生这种情况的原因可能是多样的，比如经济政策的改变、经济结构的变化，或改变某一具体行业的一项发明等。如果这样的变化或“突变”发生，那么忽视这些变化的回归模型可能会对推断和预测提供误导性偏差。

① 关于经济时间序列变量中随机性趋势及其给回归分析带来的问题的进一步讨论，请见 Stock 与 Watson(1988)。

对于一个随时间变化的时间序列回归函数来说，本节提出诊断突变的两个策略。第一个策略从假设检验方面寻找潜在的突变，并需要用 F 统计量检验回归系数的变化。第二个策略从预测方面寻找潜在的突变，你故意让你的样本比实际结束得早，并对你已做出的预测进行评价。当预测效果明显比预期效果差时，就可以发现存在突变。

12.7.1 什么是突变

突变可能会由于总体回归系数在某一确定日期的某种离散变化而产生，也可能会因为系数在一段较长的时间里逐渐演变而出现。

在宏观经济数据中，宏观经济政策的重要变化是离散性突变产生的一个原因。例如，1972 年，固定汇率的布雷顿森林体系的崩溃，造成了美元/英镑汇率的时间序列行为的突变，这在图 12—2(b)中可以明显地看出。在 1972 年以前，汇率基本上是不变的，除了 1968 年惟一一次英镑贬值，当时英镑相对于美元的官方价值降低了。相反，自 1972 年以来，汇率在非常广泛的范围内波动。

突变也可能会因为总体回归随时间的演变而更缓慢地发生。例如，经济政策的缓慢演变和经济结构的不断变化就可能导致这种突变缓慢地发生。本节所描述的检测突变的方法，可用来识别两类突变，即确定性变化和缓慢演变。

突变所引致的问题。如果在样本期间总体回归函数中发生了突变，那么全样本的 OLS 回归估计值会估计“在平均意义上”成立的关系，也就是说，这个估计值合并了两个不同的时期。在样本末期，“平均的”回归函数可能完全不同于真正的回归函数，这要取决于突变的位置和大小，因此突变会导致较差的预测。

12.7.2 检验突变

识别突变的一种方法是检验回归系数的离散变化或突变，具体怎样做取决于所怀疑的突变发生的日期(**突变日期(break date)**)是已知还是未知。

日期已知时突变的检验。在某些应用中，你可能会怀疑在一个已知的日期存在突变。例如，如果用 20 世纪 70 年代的数据研究国际贸易关系，你可能会假设汇率的总体回归函数在 1972 年存在突变，那时固定汇率的布雷顿森林体系被放弃，取而代之的是浮动汇率体系。

如果假设的系数突变的日期是已知的，那么没有突变这一零假设可用第 6 章中(见重要概念 6.4)所研究的二元变量交互回归进行检验。为了使问题简单，考虑一个 ADL(1,1) 模型，这样有一个截距项、一个 Y_t 的滞后项和一个 X_t 的滞后项。设 τ 表示假设的突变日期，并设 $D_t(\tau)$ 是一个二元变量，它在突变日期之前等于 0，而在突变日期之后等于 1，因此，当 $t \leq \tau$ 时，$D_t(\tau)=0$；当 $t>\tau$ 时，$D_t(\tau)=1$。所以，包括这个二元突变指示变量和所有交叉项的回归是：

$$Y_t=\beta_0+\beta_1 Y_{t-1}+\delta_1 X_{t-1}+\gamma_0 D_t(\tau)+\gamma_1[D_t(\tau)\times Y_{t-1}]+\gamma_2[D_t(\tau)\times X_{t-1}]+u_t \qquad (12.35)$$

如果不存在突变，那么两部分样本的总体回归函数应相同，因此涉及突变二元变量 $D_t(\tau)$ 的项就不会进入方程(12.35)中。也就是说，在不存在突变的零假设下，$\gamma_1=\gamma_2=\gamma_3=0$。在存在突变的备择假设下，总体回归函数在突变日期 τ 的前后是不同的，在此情况下至少有一个 γ 是非零的。因此，存在突变的假设可以在至少有一个 γ 是非零的备择假设下，通过检验零假设 $\gamma_1=\gamma_2=\gamma_3=0$ 的 F 统计量进行检验，这个检验常被称为已知突变日期的邹检验，是以其创始人 Gregory Chow(1960)的名字来命名的。

如果存在多个预测因子或更多个滞后项，那么这个检验可以通过对所有回归因子构造



二元可变的交互变量，并检验所有涉及 $D_t(\tau)$ 项的所有系数都为 0 的假设来扩展这个检验。

这种方法可以进行修改以检验系数子集中的突变，通过对所感兴趣的回归因子的子集只引入二元变量的交互项来实现。

突变日期未知时突变的检验。通常可能的突变日期是未知的，或只在一定范围内是已知的。例如，假设你怀疑突变发生在两个日期 τ_0 和 τ_1 之间的某个时候，对邹检验做适当修改后就可处理这类问题。具体来说，可通过检验在 τ_0 和 τ_1 之间所有可能日期 τ 的突变，然后用所得的最大的 F 统计量来检验未知日期的突变。这个修正的邹检验有个不同的称谓，叫做 Quandt 似然比（QLR）统计量（Quandt likelihood ratio statistic）（Quandt，1960）（我们要用这个术语），或更晦涩地称其为 sup-Wald 统计量。

由于 QLR 统计量是许多 F 统计量中最大的那个，因此，它的分布和单个 F 统计量的分布是不同的。然而，QLR 统计量的临界值必须从一个特殊的分布中得到。像 F 统计量一样，这个分布依赖于所检验的约束条件个数 q，即在备择假设下允许突变或变化的系数的个数（包括截距项）。QLR 统计量还依赖于 τ_0/T 和 τ_1/T，即依赖于用来计算 F 统计量的子样本的端点 τ_0 和 τ_1，它被表示为总样本容量的比重。

为了使 QLR 统计量分布的大样本近似做得比较好，子样本端点 τ_0 和 τ_1 不能太接近于样本端点。由于这个原因，实际中的 QLR 统计量是在样本的一个“修匀”范围或子集上计算的。通常的选择是使用 15% 修匀，即设 $\tau_0 = 0.15T$ 和 $\tau_1 = 0.85T$（四舍五入到整数）。在 15% 修匀下，突变日期的 F 统计量是在样本 70% 的中心部分计算的。

表 12—5 给出了在 15% 修匀下计算出来的 QLR 统计量的临界值。将这些临界值与 $F_{q,\infty}$ 分布的临界值（见附表 4）相比较，结果表明 QLR 统计量的临界值更大。这反映了 QLR 统计量是许多单个 F 统计量中的最大的那个。通过检查许多可能突变日期的 F 统计量，QLR 统计量有许多机会可拒绝，这导致了 QLR 临界值比单个 F 统计量的临界值大。

像邹检验一样，QLR 统计量也可被用来集中关注只在某些回归系数上存在突变的可能性。操作方法如下：首先使用只对可疑系数变量的二元变量交互项计算不同突变日期的邹检验值，然后计算在 $\tau_0 \leq \tau \leq \tau_1$ 范围内最大的那个邹检验值。这种形式的 QLR 检验的临界值也可以从表 12—5 中取得，其中约束条件的个数（q）是由所构成的 F 检验所检验的约束条件的个数。

如果检验范围内的某个日期存在离散性突变，那么在大样本条件下，QLR 检验统计量会以很高的概率拒绝零假设。此外，所构成的 F 统计量取其最大值的日期 $\hat{\tau}$ 是突变日期 τ 的一个估计值。这个估计值是个好的估计值，因为在一定的技术条件下，$\hat{\tau}/T \xrightarrow{P} \tau/T$，也就是说，在整个样本期间发生突变的比例被一致地估计了。

当存在多个离散突变，或突变以回归函数缓慢演变的形式出现时，QLR 统计量在大样本条件下也以很高的概率拒绝零假设。这意味着 QLR 统计量识别的是不稳定性的形式，而不是单个离散的突变。因此，如果 QLR 统计量拒绝了零假设，那么可能意味着存在单个离散突变，或存在多个离散突变，或存在回归函数的缓慢演变。

QLR 统计量在重要概念 12.9 中总结。



警告：你可能并不知道突变日期，即使你自己认为已经知道了。有时候，某位专家可能会认为自己已经知道了可能的突变日期，因此可以用邹检验而不用 QLR 检验。但如果这种认识是以专家对所分析的序列的认识为基础的话，那么实际上这个日期也是使用数据估计出来的，尽管是以非正式的方式。对突变日期的初步估计意味着，通常的 F 临界值不能用

于那个日期突变的邹检验。因而，在这种情况下使用 QLR 统计量仍然是适当的。

表 12—5　　15% 修匀的 QLR 统计量临界值

约束条件个数(q)	10%	5%	1%
1	7.12	8.68	12.16
2	5.00	5.86	7.78
3	4.09	4.71	6.02
4	3.59	4.09	5.12
5	3.26	3.66	4.53
6	3.02	3.37	4.12
7	2.84	3.15	3.82
8	2.69	2.98	3.57
9	2.58	2.84	3.38
10	2.48	2.71	3.23
11	2.40	2.62	3.09
12	2.33	2.54	2.97
13	2.27	2.46	2.87
14	2.21	2.40	2.78
15	2.16	2.34	2.71
16	2.12	2.29	2.64
17	2.08	2.25	2.58
18	2.05	2.20	2.53
19	2.01	2.17	2.48
20	1.99	2.13	2.43

注：当 $\tau_0 = 0.15T, \tau_1 = 0.85T$（四舍五入到整数）时，这些临界值适用，因此，$F$ 统计量是对样本 70% 的中心部分中所有潜在的突变日期计算的。约束条件个数 q 是每个单个 F 统计量所检验的约束条件的个数。很感谢 Donald Andrwes 向我们提供这个表格，取代了 Andrwes(1993) 中的表 1。



重要概念 12.9

系数稳定性的 QLR 检验

设 $F(\tau)$ 表示在时期 τ 的回归系数中存在突变这个假设的 F 统计量，例如，在公式(12.35)的回归中，就是检验零假设 $\gamma_1 = \gamma_2 = \gamma_3 = 0$ 的 F 统计量。QLR(或 sup-Wald)检验是在 $\tau_0 \leqslant \tau \leqslant \tau_1$ 范围内的最大的那个统计量。

$$QLR = \max[F(\tau_0), F(\tau_0 + 1), \cdots, F(\tau_1)] \tag{12.36}$$

1. 和 F 统计量一样，QLR 统计量可用来检验全部回归系数的突变，也可用来检验只有部分回归系数的突变。

2. 在大样本条件下，QLR 统计量的分布在零假设下依赖于所检验的约束条件个数 q 和作为 T 的一个部分的端点 τ_0 和 τ_1。15% 修匀的临界值($\tau_0 = 0.15T$ 和 $\tau_1 = 0.85T$，四舍五入

到整数)在表 12—5 中给出。

3. QLR 检验可识别单个离散突变、多个离散突变和回归函数的缓慢演变。

4. 如果回归函数中存在一个确定的突变,那么最大的邹统计量发生的日期是突变日时期的一个估计量。

应用:菲利浦斯曲线稳定吗? QLR 检验提供了一种检验从 1962 年到 1999 年菲利浦斯曲线是否稳定的方法。具体来说,我们关注在 ΔInf_t 和 $Unemp_t$ 每个都包含四阶滞后项的公式(12.17)的 ADL(4,4)设定中,失业率的滞后值系数和截距项是否已发生了变化。

公式(12.17)中的截距项和 $Unemp_{t-1},\cdots,Unemp_{t-4}$ 的系数都为常数的零假设和在给定日期发生突变的备择假设,检验这两个假设的邹检验 F 统计量的图形在图 12—5 中绘制,图中给出了 70% 的样本中心部分的突变。例如,检验在 1980: Ⅰ 存在突变的 F 统计量为 2.26,该日期的值在图 12—5 中做了标示。每个 F 统计量检验 5 个约束条件(截距项和失业率滞后项的 4 个系数都不发生变化),因此,$q=5$。这些 F 统计量的最大值是 3.53,它发生在 1982: Ⅱ,这就是 QLR 统计量。将 3.53 与表 12.5 中 $q=5$ 的临界值相比较表明,"这些系数是稳定的"假设在 10% 的显著性水平下(临界值是 3.26)被拒绝,但在 5% 的显著性水平下(临界值是 3.66)不能被拒绝。因而,存在一定的证据表明,这 5 个系数中至少有一个系数在样本期间已发生变化,但这种证据并不特别明显。



注:在给定的突变日期,这里所绘制的 F 统计量检验方程(12.17)中的 $Unemp_{t-1}$, $Unemp_{t-2}$, $Unemp_{t-3}$, $Unemp_{t-4}$ 系数或截距中至少一个有突变的零假设。例如,检验在 1980: Ⅰ 突变的 F 统计量为 2.26。QLR 统计量是这些 F 统计量中的最大值,它是 3.53,它超过 10% 的临界值 3.26,但小于 5% 的临界值 3.66。

图 12.5 在不同日期检验方程(12.17)中突变的 F 统计量



12.7.3 伪样本外预测

一个模型被估计出来之后,对该预测模型的最终检验是它在样本外的表现效果,也即它在"真实世界"里的预测效果。伪样本外预测(pseudo out-of-sample forecasting)是对一个预测模型模拟其在真实时间表现效果的方法。伪样本外预测的思想很简单:选择一个在样

本终点附近的日期,用直到该日期的数据估计你的预测模型,然后用那个所估计的模型进行预测。对样本终点附近的多个日期执行这个操作,得到一系列的伪预测值和伪预测误差。于是,这个伪预测误差就能被用来判断,如果预测关系是平稳的,那么这些预测是否代表了你所期望的预测值。

之所以称其为"伪"样本外预测,是因为它并不是真实的样本外预测。真实的样本外预测发生在真实的世界里,也就是说,在不知道一个序列未来值给你所带来的好处的条件下进行预测。在伪样本外预测中,你用你的模型模拟真实时间的预测,但你有"未来的"数据,依靠它们评价那些模拟的预测或伪预测。伪样本外预测模拟了在真实世界里可能发生的预测过程,但不用等待新数据的到来。

伪样本外预测给预测者一种在样本末端模型被预测得好坏的感觉。这可以提供有价值的信息,要么支持该模型一直被预测得很好这种信心,要么表明该模型在刚刚过去的时间里远离了轨道。伪样本外预测方法在重要概念 12.10 中总结。

重要概念 12.10

伪样本外预测

伪样本外预测按下列步骤计算:

1. 选择一个用于生成伪样本外预测的观测期数 P,例如,P 可能是样本容量的 10% 或 15%,设 $s = T - P$。

2. 用缩短了的这个数据集对 $t = 1,\cdots,s$ 估计预测的回归方程。

3. 计算在这个缩短的样本之外第一个时期 $s+1$ 的预测值,我们将其称之为 $\widetilde{Y}_{s+1|s}$。

4. 计算预测误差,$\widetilde{u}_{s+1} = Y_{s+1} - \widetilde{Y}_{s+1|s}$。

5. 对余下的时期,从 $s = T-P+1$ 到 $T-1$,重复步骤 2~4(在每个时期重新估计这个回归方程)。伪样本外预测值是 $\{\widetilde{Y}_{s+1|s}, s = T-P,\cdots,T-1\}$,伪样本外预测误差是 $\{\widetilde{u}_{s+1}, s = T-P,\cdots,T-1\}$。

伪样本外预测的其他用途。伪样本外预测的第二个用途是估计 RMSFE。由于伪样本外预测值只用预测时期以前的数据计算,因此,伪样本外预测误差反映了与误差项的未来值有关的不确定性和由于估计回归系数所引起的不确定性,也就是说,这个伪样本外预测值包含了公式(12.21)中的两个误差来源。因而,这个伪样本外预测误差的样本标准差是 RMSFE 的一个估计量。如 12.4 节中所讨论的,RMSFE 的这个估计量能用来量化预测不确定性和构造预测区间。

伪样本外预测的第三个用途是,对两个或更多的候选预测模型进行比较。两个看起来同样好地拟合了样本数据的模型,在伪样本外预测练习中的表现可能完全不同。当模型不同时,例如,当它们包含不同的预测因子时,伪样本外预测提供了比较这两个模型是否能提供可靠预测结果的一种简便方法。

应用:菲利浦斯曲线在 20 世纪 90 年代是否发生了变化?如果菲利浦斯曲线的系数在 20 世纪 90 年代发生变化,那么在此期间计算的伪样本外预测值会变得很糟糕。从 1994:Ⅰ到 1999:Ⅳ期间,用四阶滞后的菲利浦斯曲线所计算的通货膨胀的伪样本外预测值,连同通货膨胀的实际值一起绘制在图 12—6 中。例如,1994:Ⅰ时期通货膨胀的预测值是这样来计算的:使用从开始到 1993:Ⅳ的数据,用 ΔInf_t 对 $\Delta Inf_{t-1},\cdots,\Delta Inf_{t-4}, Unemp_{t-1},\cdots,Unemp_{t-4}$

而且这种不稳定性导致了偏高的通货膨胀变化的预测值。在这个模型用于真实世界的预测之前，尽力识别这个漂移的来源，并将其综合到菲利浦斯曲线模型的修正形式中，这一点是非常重要的。

放到一起来说，这种伪样本外预测中的偏差和 QLR 统计量（在 10% 的显著性水平下）对平稳性的拒绝，表明了这个四阶滞后的菲利浦斯曲线是不平稳的。这种不平稳性在 20 世纪 90 年代期间和 21 世纪初期是很值得考虑的问题，因为经济预测者认识到基于菲利浦斯曲线的通货膨胀预测值太高，如图 12—6 所示。一些宏观经济学家认为这种不平稳性的根源是 20 世纪 90 年代自然失业率的下降，它使我们所研究的回归的截距向负的方面变化。不过，其他一些宏观经济学家却认为这种分析是更完全的，菲利浦斯曲线的整体概念——连接过度需求压力和整体价格通货膨胀之间的关系——只是前信息时代经济的一个陈旧的特征。如果有兴趣进一步阅读关于这方面的争论，请见《经济展望杂志》1997 年冬季刊关于菲利浦斯曲线的专题讨论。

12.7.4 避免由突变所引致的问题

调整总体回归函数中突变的最好方法取决于该突变的来源。如果一个确定的突变发生在某一特定的日期，那么这个突变将会以很高的概率被 QLR 统计量所识别，而且能够估计出突变发生的日期。因此，可以使用代表与这个突变有关的两个子样本的一个二元变量来估计这个回归函数，必要时可引入其他的回归因子起交互作用。如果所有的系数都发生了突变，那么这个回归采纳公式（12.35）的形式，其中 τ 由所估计的突变日期 $\hat{\tau}$ 代替，但如果只有部分系数发生了突变，那么只有相关的交互作用项出现在回归中。如果确实存在确定的突变，那么关于回归系数的推断可以如往常一样进行，比如使用基于 t 统计量的假设检验通常的正态临界值。此外，还可使用适用于样本末端的所估计的回归函数进行预测。

如果突变不明显，而是由参数缓慢的正在进行的变化所引起的，那么补救措施更为困难，而且超出了本书的范围[①]。

12.8 结论

在时间序列数据中，变量在不同观测值或时期之间通常是相关的。这种相关的结果是，可以根据该时间序列的当前值和过去值，使用线性回归方法预测它的未来值。时间序列回归的出发点是自回归，也即回归因子是因变量滞后值的回归。如果可获得额外的预测因子，那么可以将它们的滞后项添加到回归中。

本章考虑了在使用时间序列数据估计和使用回归时出现的若干技术问题。一个问题是确定包含在回归中的滞后项的个数。如 12.5 节中所讨论的，如果选择的是使 BIC 最小化的滞后阶数，那么所估计的滞后长度与真实的滞后长度是一致的。

另一个问题是，所分析的序列是不是平稳的。如果所分析的序列是平稳的，那么就可使用通常的统计推断方法（例如，将 t 统计量与正态分布临界值进行比较），而且因为总体回归函数随时间推移是稳定的，所以用历史数据所估计的回归方程能被可靠地用于预测。但如果所研究的序列是非平稳的，那么情况就会变得更加复杂，其中具体的复杂程度取决于这个非平稳性的性质。例如，如果一个序列因含有随机性趋势而变成是非平稳的，那么所得的

① 对存在离散性突变时的估计和检验的更多讨论，请见 Hansen（2001）。对存在系数缓慢演变时突变的估计和预测的高级讨论，请见 Hamilton（1994，第 13 章）。

迪基—富勒(ADF)统计量　突变日期　Quandt 似然比(QLR)统计量　伪样本外预测

复习概念

12.1　观察图 12—2(c)中日本实际 GDP 的对数图。这个时间序列看上去是平稳的吗？请解释说明。假设你计算了这个序列的一阶差分，它看上去会是平稳的吗？请解释说明。

12.2　许多金融经济学家认为随机游动模型是股票价格对数的一个很好的描述。这隐含着，股票价格的百分比变化是不可预测的。一位金融分析师声称他有一个比随机游动模型预测能力更好的新模型。请解释你将如何检验这位分析师认为他的模型更好这一观点。

12.3　一位研究人员估计了一个含有截距的 AR(1)，并求出 β_1 的 OLS 估计值为 0.95，标准误为 0.02。一个 95% 的置信区间会包含 $\beta_1 = 1$ 吗？请解释说明。

12.4　假设你怀疑公式(12.17)中的截距在 1992:I 发生了变化，你打算如何修正这个方程以体现这个变化？你怎样检验该截距的变化？如果你不知道变化的日期，你将如何检验该截距的变化？

练习

*12.1　假设 Y_t 满足平稳的 AR(1)模型 $Y_t = 2.5 + 0.7Y_{t-1} + u_t$，其中，$u_t$ 为独立同分布，$E(u_t) = 0$，$\text{var}(u_t) = 9$。

a. 计算 Y_t 的均值和方差。

b. 计算 Y_t 的前二阶自协方差。

c. 计算 Y_t 的前二阶自相关系数。

d. 假设 $Y_T = 102.3$，计算 $Y_{T+1|T} = E(Y_{T+1} | Y_T, Y_{T-1}, \cdots)$。

12.2　工业生产指数(IP_t)是一个测量某一月份内生产的工业产品数量的月度时间序列。本问题使用美国关于这个指数的数据。所有的回归都是在 1960:1 到 2000:12(即从 1960 年 1 月份到 2000 年 12 月份)的样本期间估计的。设 $Y_t = 1\,200 \times \ln(IP_t/IP_{t-1})$。

a. 预测者说，用年度百分点测度的 Y_t 反映了 IP 月度百分比的变化。他的观点正确吗？为什么？

b. 假设一位预测者对 Y_t 估计了如下的 AR(4)模型：

$$\hat{Y}_t = \underset{(0.062)}{1.377} + \underset{(0.078)}{0.318}\ Y_{t-1} + \underset{(0.055)}{0.123}\ Y_{t-2} + \underset{(0.068)}{0.068}\ Y_{t-3} + \underset{(0.056)}{0.001}\ Y_{t-4}$$

利用表 12—6 中从 2000 年 7 月到 2000 年 12 月的 IP 值，使用这个 AR(4)模型预测 2001 年 1 月的 Y_t 值。

表 12—6　**2000 年 7 月到 2000 年 12 月的 IP 值**

时期	2000:7	2000:8	2000:9	2000:10	2000:11	2000:12
IP	147.595	148.650	148.973	148.660	148.206	147.300

c. 由于担心生产中存在潜在的季节波动，该预测者将 Y_{t-12} 添加到了自回归模型中。所估计的 Y_{t-12} 的系数是 −0.054，标准误为 0.053。这个系数在统计上是否是显著的？

d. 由于担心存在潜在的突变，她计算了 AR(4)模型中的常数项和 AR 系数的(在 15% 修匀下)QLR 检验统计量，相应的 QLR 统计量为 3.45。这里存在突变的证据吗？请解释

说明。

e. 由于担心自己可能在模型中包含了太多或者太少的滞后项，该预测者在同样的样本期间估计了 $p=1,\cdots,6$ 的 AR(p) 模型。这些估计模型中每一个模型的残差平方和都显示在表 12—7 中。用 BIC 准则估计自回归中应该包含的滞后阶数。如果使用 AIC，结果会不同吗？

表 12—7　　**估计模型的残差平方和**

AR 阶数	1	2	3	4	5	6
SSR	29 175	28 538	28 393	28 391	28 378	28 317

*12.3　利用与练习 12.2 相同的数据，一位研究人员用下列回归检验了 $\ln(IP_t)$ 中的随机性趋势：

$$\widehat{\Delta\ln(IP_t)} = \underset{(0.024)}{0.061} + \underset{(0.00001)}{0.00004t} - \underset{(0.007)}{0.018}\ \ln(IP_{t-1}) + \underset{(0.075)}{0.333}\ \Delta\ln(IP_{t-1}) + \underset{(0.055)}{0.162}\ \Delta\ln(IP_{t-2})$$

其中，显示在括号内的标准误是用仅适用于同方差的公式计算出来的，而且回归因子“t”是一个线性时间趋势。

a. 用 ADF 统计量检验 $\ln(IP)$ 中的随机性趋势（单位根）。

b. 这些结论是否支持练习 12.2 中所使用的设定？请解释说明。

12.4　练习 12.2 中的那位预测者将她对 IP 增长的 AR(4) 模型扩展为包含 ΔR_t 的四阶滞后值，这里 R_t 是 3 个月美国短期国债利率（用年度百分点测量）。

a. ΔR_t 的四阶滞后的格兰杰因果关系 F 统计量为 2.35。利率有助于预测 IP 的增长吗？请解释说明。

b. 该研究人员还用 ΔR_t 对一个常数、ΔR_t 的四阶滞后和 IP 增长的四阶滞后进行回归，所得的关于 IP 增长的四阶滞后的格兰杰因果关系 F 统计量为 2.87。IP 增长有助于预测利率吗？请解释说明。

12.5　证明下列关于条件均值、预测值和预测误差的结论。

a. 设 W 是个均值为 μ_W 且方差为 σ_W^2 的随机变量，并设 c 为一常数。证明：$E[(W-c)^2]=\sigma_W^2+(\mu_W-c)^2$。

b. 利用关于 $Y_{t-1},Y_{t-1},\cdots$ 的数据，考虑预测 Y_t 的问题。设 f_{t-1} 表示 Y_t 的某个预测值，其中 f_{t-1} 的下角标 $t-1$ 表示预测值是从开始到 $t-1$ 期数据的函数，设 $E[(Y_t-f_{t-1})^2|Y_{t-1},Y_{t-2},\cdots]$ 是预测值 f_{t-1} 的条件均方误差，它是以从开始到 $t-1$ 期 Y 的观察值为条件的。证明：当 $f_{t-1}=Y_{t|t-1}$ 时，条件均方预测误差最小，这里 $Y_{t|t-1}=E(Y_t|Y_{t-1},Y_{t-2},\cdots)$。（提示：将 (a) 的结论推广到条件期望）

c. 证明：AR(p)（重要概念 12.3 中的公式 (12.14)）的误差项 u_t 是序列无关的。（提示：利用公式 (2.25)）

12.6　在本练习中，你将执行 12.6 节中所讨论的一个研究伪回归现象的蒙特卡洛模拟实验。在蒙特卡洛研究中，用计算机生成人工数据，然后用这些人工数据计算所研究的统计量。当已知模型统计量分布的数学表达式很复杂（像这里一样）甚至是未知的时，蒙特卡洛模拟实验使得计算这些统计量的分布成为可能。在本练习中，你将生成数据，使得两个序列 Y_t 和 X_t 是独立分布的随机游动，具体步骤是：

ⅰ. 用计算机生成一列 $T=100$ 的独立同分布的标准正态随机变量。将这些变量记为 $e_1,e_2,\cdots,e_{100}$。设 $Y_1=e_1$，$Y_t=Y_{t-1}+e_t$，$t=2,3,\cdots,100$。



ⅱ. 用计算机生成一个新的、$T=100$ 的独立同分布标准正态随机变量 $a_1,a_2,\cdots,a_{100}$。设 $X_1=a_1, X_t=X_{t-1}+a_t, t=2,3,\cdots,100$。

ⅲ. 用 Y_t 对常数和 X_t 回归。计算 OLS 估计量、回归的 R^2，以及检验 β_1（X_t 的系数）为 0 的零假设的（仅适用于同方差的）t 统计量。

根据以上算法回答下列问题：

a. 从（ⅰ）-（ⅲ）进行一次运算。利用通常的5%临界值1.96，使用（ⅲ）中的 t 统计量检验 $\beta_1=0$ 的零假设。你所估计的回归的 R^2 是多少？

b. 重复（a）1 000 次，保存每个 R^2 和 t 统计量的值。构造一个 R^2 和 t 统计量的直方图。R^2 和 t 统计量分布的5%，50%，95%百分位数分别是多少？在1 000个模拟数据集中，t 统计量的绝对值大于1.96的比重有多大？

c. 按不同的观测值数重复（b）的步骤，例如 $T=50$ 和 $T=200$。随着样本容量的增加，由于已生成的 Y 和 X 是独立分布的，拒绝零假设的次数比重接近5%吗？随着 T 变大，这个比重值看起来会接近于某个其他的极限值吗？那个极限值是多少？

附录12.1 第12章中所使用的时间序列数据

美国的宏观经济时间序列数据由不同的政府机构搜集和发布。美国的消费者价格指数是用月度调查数据测度的，并由劳工统计局（BLS）编辑。失业率是根据 BLS 的当前人口普查（见附录3.1）的数据计算的。这里所用的季度数据是根据对月度数值取平均数计算的。联邦基金利率是联邦储备委员会报告的日利率的月度平均，而美元/英镑汇率数据是日汇率的月度平均，两者都是对该季度的最后月份而言的。日本的实际 GDP 数据是从 OECD 得到的。NYSE 综合指数的日百分比变化是用 $100\Delta\ln(NYSE_t)$ 计算的，其中 $NYSE_t$ 是纽约股票交易所的日收盘指数值。因为股票交易所在周末和节假日不营业，所以，分析的时期是营业日。这些以及许多其他的经济时间序列在不同数据搜集机构的相关网站上可以免费获得。



附录12.2 AR(1)模型中的平稳性

本附录证明了，如果 $|\beta_1|<1$ 且 u_t 是平稳的，那么 Y_t 就是平稳的。回想一下重要概念12.5中的内容，如果 $(Y_{s+1},\cdots,Y_{s+T})$ 的联合分布不依赖于 s，那么时间序列变量 Y_t 就是平稳的。为了使论证简单化，在简化的假设 $\beta_0=0$ 和 $\{u_t\}$ 服从独立同分布 $N(0,\sigma_u^2)$ 下，我们正式证明 $T=2$ 的情形。

第一步是依据 u_t 推导 Y_t 的表达式。由于 $\beta_0=0$，因此公式（12.8）隐含着 $Y_t=\beta_1 T_{t-1}+u_t$，将 $Y_{t-1}=\beta_1 Y_{t-2}+u_{t-1}$ 代入到这个表达式中，得到 $Y_t=\beta_1(\beta_1 Y_{t-2}+u_{t-1})+u_t=\beta_1^2 Y_{t-2}+\beta_1 u_{t-1}+u_t$。继续另一步的这个替代，得到 $Y_t=\beta_1^3 Y_{t-3}+\beta_1^2 u_{t-2}+\beta_1 u_{t-1}+u_t$，无限期地继续下去，可得到：

$$Y_t=u_t+\beta_1 u_{t-1}+\beta_1^2 u_{t-2}+\beta_1^3 u_{t-3}+\cdots=\sum_{i=0}^{\infty}\beta_1^i u_{t-i} \tag{12.37}$$

因而，Y_t 是 u_t 的当前值和过去值的一个加权平均。因为 u_t 服从正态分布，而且正态随机变量的加权平均也是正态的（见2.6节），所以 Y_{s+1} 和 Y_{s+2} 服从双变量的正态分布。回想一下2.6节，双变量的正态分布完全由两个变量的均值、方差和协方差所决定。因此，要证明 Y_s 是平稳的，我们需要证明 (Y_{s+1},Y_{s+2}) 的均值、方差和协方差不依赖于 s。下面使用的该

命题的扩展形式可用来证明$(Y_{s+1},\cdots,Y_{s+T})$的分布不依赖于s。

Y_{s+1}和Y_{s+2}的均值和方差可用公式(12.37)来计算,但需要用下角标$s+1$或$s+2$代替t。首先,由于对所有的t有$E(u_t)=0$,$E(Y_t)=E(\sum_{i=0}^{\infty}\beta_1^i u_{t-i})=\sum_{i=0}^{\infty}\beta_1^i E(u_{t-i})=0$,因此$Y_{s+1}$和$Y_{s+2}$的均值都为0,尤其是它们不依赖于$s$。其次,$\text{var}(Y_t)=\text{var}(\sum_{i=0}^{\infty}\beta_1^i u_{t-i})=\sum_{i=0}^{\infty}(\beta_1^i)^2\text{var}(u_{t-i})=\sigma_u^2\sum_{i=0}^{\infty}(\beta_1^i)^2=\sigma_u^2/(1-\beta_1^2)$,这里最后一个等式由下列事实得出、如果$|\alpha|<1$,则$\sum_{i=0}^{\infty}\alpha^i=1/(1-\alpha)$,因而$\text{var}(Y_{s+1})=\text{var}(Y_{s+2})=\sigma_u^2/(1-\beta_1^2)$,只要$|\beta_1|<1$,它就不依赖于$s$。最后,由于,$Y_{s+2}=\beta_1 Y_{s+1}+u_{s+2}$,$\text{cov}(Y_{s+1},Y_{s+2})=E(Y_{s+1},Y_{s+2})=E[Y_{s+1}(\beta_1 Y_{s+1}+u_{s+2})]=\beta_1\text{var}(Y_{s+1})+\text{cov}(Y_{s+1},u_{s+2})=\beta_1\text{var}(Y_{s+1})=\beta_1\sigma_u^2/(1-\beta_1^2)$,因此,协方差不依赖于$s$。所以,$Y_{s+1}$和$Y_{s+2}$具有一个不依赖于$s$的联合概率分布,也就是说,它们的联合分布是平稳的。如果$|\beta_1|\geq 1$,那么这个计算就无效,因为公式(12.37)中的无穷和不收敛,Y_t的方差是无限的。因此,如果$|\beta_1|<1$,那么Y_t就是平稳的,但如果$\beta_1=1$,则Y_t是非平稳的。

前面的争论是在$\beta_0=0$和u_t服从正态分布的假设下进行的。如果$\beta_0\neq 0$,那么结论是相似的,但Y_{s+1}和Y_{s+2}的均值变成了$\beta_0/(1-\beta_1)$,而且必须对方程(12.37)进行非零均值的修正。u_t是独立同分布的正态假设,可被u_t是平稳的并且具有有限的方差这一假设来代替,因为依据公式(12.37),Y_t仍可以表示为u_t的当前值以及过去值的函数,所以只要u_t的分布是平稳的,且公式(12.37)中的无穷和是有意义的(也即它收敛,这要求$|\beta_1|<1$),那么Y_t的分布就是平稳的。

附录 12.3　滞后算子符号

通过采用熟知的滞后算子符号,本章以及下两章中的符号被大大地简化。设L表示滞后算子(lag operator),它具有将变量变换为它的滞后项的性质,也就是说,滞后算子L具有性质$LY_t=Y_{t-1}$。通过使用两次滞后算子,可以得到二阶滞后,即$L^2Y_t=L(LY_t)=LY_{t-1}=Y_{t-2}$。更一般地说,通过使用$j$次滞后算子,可以得到$j$阶滞后。总之,滞后算子具有如下性质:

$$LY_t=Y_{t-1},\ L^2Y_t=Y_{t-2},\ L^jY_t=Y_{t-j} \tag{12.38}$$

滞后算子符号允许我们定义**滞后多项式**(lag polynomial),它是一个用滞后算子表示的多项式。

$$\alpha(L)=\alpha_0+\alpha_1 L+\alpha_2 L^2+\cdots+\alpha_p L^p=\sum_{j=0}^{p}\alpha_j L^j \tag{12.39}$$

其中,$\alpha_0,\cdots,\alpha_p$是滞后多项式的系数,且$L^0=1$。公式(12.39)中滞后多项式$\alpha(L)$的次数为p。用$\alpha(L)$乘以Y_t得到:

$$\alpha(L)Y_t=(\sum_{j=0}^{p}\alpha_j L^j)Y_t=\sum_{j=0}^{p}\alpha_j(L^jY_t)=\sum_{j=0}^{p}a_jY_{t-j}=\alpha_0Y_t+\alpha_1Y_{t-1}+\cdots+\alpha_pY_{t-p} \tag{12.40}$$

公式(12.40)中的表达式隐含着公式(12.14)中的AR(p)模型可被简写为:

$$\alpha(L)Y_t=\beta_0+u_t \tag{12.41}$$

其中,$\alpha_0=1$且$\alpha_j=-\beta_j$,$j=1,\cdots,p$。同样,ADL(p,q)模型可写为:

$$\alpha(L)Y_t = \beta_0 + c(L)X_{t-1} + u_t \qquad (12.42)$$

其中，$\alpha(L)$是一个次数为$p(a_0=1)$的滞后多项式，$c(L)$是一个次数为$q-1$的滞后多项式。

附录 12.4 ARMA 模型

自回归—移动平均模型(auto regressive-moving average model，简写为 ARMA)通过将u_t建立为序列相关模型，具体来说是将u_t建立为另一个不可观测的误差项的分布滞后(或"移动平均")模型，从而推广了自回归模型。也就是说，用附录 12.3 的滞后算子符号表示，设$u_t = b(L)e_t$，其中e_t是一个序列无关的不可观测的随机变量，而$b(L)$是一个$b_0=1$、次数为q的滞后多项式。那么，ARMA(p,q)模型为：

$$\alpha(L)Y_t = \beta_0 + b(L)e_t \qquad (12.43)$$

其中，$\alpha(L)$是一个$a_0=1$、次数为p的滞后多项式。

AR 和 ARMA 模型都可被看做是近似得出Y_t自协方差的方法，其原因在于，任何具有有限方差的平稳时间序列Y_t，都可被写成一个带有序列无关误差项的 AR 或 MA，尽管 AR 或 MA 模型可能需要含有无穷阶数。第二个结果，一个平稳过程可被写成移动平均的形式，就是著名的沃尔德(Wold)分解理论，它是支撑平稳时间序列分析理论的基本结论之一。

理论上，只要滞后多项式有足够高的阶数，AR，MA 和 ARMA 模型族便具有相同丰富的意义。在某些情况下，用含有较小的p和q的 ARMA(p,q)模型近似替代自协方差，比仅用少数滞后项的纯粹的 AR 模型要好。然而实际上，ARMA 模型的估计要比 AR 模型的估计更困难，而且 ARMA 模型比 AR 模型更难被扩展到包含额外的回归因子。

附录 12.5 BIC 滞后长度估计量的一致性

本附录概述了"自回归模型中滞后长度的 BIC 估计量$\hat{p}$在大样本条件下是正确的"这一论述，即$\Pr(\hat{p}=p)\to 1$，但这不适合 AIC 估计量，它高估了p，即使在大样本条件下。

BIC

首先考虑特殊的情形，即当真实的滞后长度为 1 时，在含有零阶、一阶或二阶滞后的自回归模型中用 BIC 选择阶数。下面证明：(ⅰ) $\Pr(\hat{p}=0)\to 0$，(ⅱ) $\Pr(\hat{p}=2)\to 0$，由此得出$\Pr(\hat{p}=p)\to 1$。将这个论证推广到$0\leq p\leq p_{max}$内进行搜索的一般情形时，需要证明$\Pr(\hat{p}<p)\to 0$和$\Pr(\hat{p}>p)\to 0$。证明这些论证的方法与下面用来证明(ⅰ)和(ⅱ)所使用的方法是相同的。

证明(ⅰ)和(ⅱ)

证明(ⅰ)。要选择$\hat{p}=0$，必须满足条件 BIC(0) < BIC(1)，即 BIC(0) − BIC(1) < 0。现在 $\text{BIC}(0)-\text{BIC}(1)=\{\ln[SSR(0)/T]+(\ln T)/T\}-\{\ln[SSR(1)/T]+2(\ln T)/T\}=\ln[SSR(0)/T]-\ln[SSR(1)/T]-(\ln T)/T$。现有 $SSR(0)/T=[(T-1)/T]s_Y^2 \xrightarrow{p} \sigma_Y^2$，$SSR(1)/T \xrightarrow{p} \sigma_u^2$，$(\ln T)/T \longrightarrow 0$。那么将这两部分合并起来，就有 $\text{BIC}(0)-\text{BIC}(1) \xrightarrow{p} \ln\sigma_Y^2-\ln\sigma_u^2>0$，因为$\sigma_Y^2>\sigma_u^2$。由此得出，$\Pr[\text{BIC}(0)<\text{BIC}(1)]\to 0$，所以$\Pr(\hat{p}=0)\to 0$。

证明(ⅱ)。要选择$\hat{p}=2$，必须满足条件 BIC(2) < BIC(1) 或 BIC(2) − BIC(1) < 0。既



然 $T[\text{BIC}(2)-\text{BIC}(1)] = T\{[\ln(SSR(2)/T)+3(\ln T)/T]-[\ln(SSR(1)/T)+2(\ln T)/T]\} = T\ln[SSR(2)/SSR(1)]+\ln T = -T\ln[1+F/(T-2)]+\ln T$，其中，$F=[SSR(1)-SSR(2)]/[SSR(2)/(T-2)]$是检验 AR(2) 中 $\beta_2=0$ 这一零假设的"经验规则"F 统计量（见附录5.3）。如果 u_t 是同方差的，那么 F 服从渐近的χ_1^2 分布；否则，它服从某个其他形式的渐近分布。因而，$\Pr[\text{BIC}(2)-\text{BIC}(1)<0]=\Pr\{T[\text{BIC}(2)-\text{BIC}(1)]<0\}=\Pr\{-T\ln[1+F/(T-2)]+(\ln T)<0\}=\Pr\{T\ln[1+F/(T-2)]>\ln T\}$。随着 T 的增加，$T\ln[1+F/(T-2)]-F\to 0$（它是对数近似 $\ln(1+a)\approx a$ 的一个结果，当 $a\to 0$ 时变得更精确）。因此，$\Pr[\text{BIC}(2)-\text{BIC}(1)<0]\to\Pr(F>\ln T)\to 0$，这样便有 $\Pr(\hat{p}=2)\to 0$。

AIC

在 AR(1) 的特殊情况下，当只考虑零阶、一阶或二阶滞后时，将（ⅰ）应用于 AIC，其中用 2 代替 $\ln T$ 项，所以有 $\Pr(\hat{p}=0)\to 0$ 。将证明 BIC（ⅱ）中的全部步骤也应用于 AIC，并用 2 代替 $\ln T$ 做一下修正，因而，$\Pr[\text{AIC}(2)-\text{AIC}(1)<0]\to\Pr(F>2)>0$。如果 u_t 是同方差的，那么 $\Pr(F>2)\to\Pr(\chi_1^2>2)=0.16$，这样 $\Pr(\hat{p}=2)\to 0.16$。一般地说，当用 AIC 选择 $\hat{p}$ 时，$\Pr(\hat{p}<p)\to 0$，但 $\Pr(\hat{p}>p)$ 趋向于某个正数，所以，$\Pr(\hat{p}=p)$ 不会趋向于 1。

第13章 动态因果效应的估计

在1983年的电影《交易场》中，由Dan Aykroyd和Eddie Murphy扮演的演员利用佛罗里达州橙子如何安全地度过寒冬的内部信息在浓缩橙汁期货市场(即以一定的价格大量买卖在未来时期执行的浓缩橙汁合约的市场)上获取了上百万美元的暴利。在现实生活中，橙汁期货的交易者确实密切关注佛罗里达州的天气，佛罗里达州的霜冻冻死了佛罗里达州的橙子，而这些橙子几乎是美国制造冷冻浓缩橙汁的全部原料，因此橙汁的供给下降，价格上涨。但当佛罗里达州天气恶化时，价格确切地上涨多少？价格是立刻上涨，还是会延期？如果是延期的话，会延期多久？这些问题都是那些在橙汁期货中的实际交易者要想取得成功必须回答的。

本章研究X在现在和未来的变化对Y的效应的估计问题，即X的变化对Y的**动态因果效应**(dynamic casual effect)。例如，佛罗里达州霜冻期间对橙汁价格随时间变化的路径的效应是什么？建模和估计动态因果效应的出发点就是所谓的分布滞后回归模型，其中，Y_t被表示为X_t的当前值及过去值的函数。13.1节通过引入随时间变化的佛罗里达州寒冷天气对浓缩橙汁价格的效应的例子，介绍了分布滞后模型。13.2节进一步考察了动态因果效应的精确含义。

估计动态因果效应的一种方法，就是用OLS估计分布滞后回归模型的系数。如13.3节中所论述的，给定X的当前值和过去值(一个被称为外生性的条件，见第10章)，如果回归误差项有条件零均值，那么这个估计量就是一致的。由于被遗漏的那些决定Y_t的因素随时间的变化而相关(即由于它们是序列相关的)，因此，分布滞后模型中的误差项可能是序列相关的。这种可能性反而需要一个新的标准误的公式，即"异方差—自相关—一致性"(HAC)(heteroskedasticity-and autocorrelation-consistent)标准误公式，这是13.4节要讨论的问题。

另一种估计动态因果效应的方法(在13.5节中论述)，是将误差项中的序列相关建模为一个自回归模型，然后用这个自回归模型去推导自回归分布滞后(ADL)模型。另外，可用广义最小二乘(GLS)估计初始的分布滞后模型的系数。不过，ADL和GLS方法要求一种比迄今为止我们所用的方法更强的外生性形式：严外生性。在严外生性条件下，给定X的

过去、现在和未来值，回归误差项的条件均值为零。

13.6 节提供了对橙汁价格和天气之间关系的更全面的分析。在这个应用中，天气超出了人力控制的范围，因而是外生的（尽管如13.6 节所述，经济理论认为它不一定是严外生的）。由于外生性是估计动态因果效应的必要条件，因此，13.7 节用几个取自宏观经济学和金融学中的例子研究了这个假设。

本章建立在12.1 节到12.4 节内容的基础之上，但除了13.6 节的实证分析部分外，不需要12.5 节到12.8 节的内容。

13.1 对橙汁数据的初步考察

奥兰多，佛罗里达州橙子种植区的中心，在正常情况下是晴朗和温暖的，但偶尔会有寒潮。如果气温下降到冰点以下太久，那么许多橙子会从树上掉下来，而且如果霜冻严重的话，橙子树也会受冻。在霜冻之后，浓缩橙汁的供给下降，价格上涨，但价格上涨的时机是相当复杂的。浓缩橙汁是“耐用的”或可储藏的商品，即它可以在冻结状态下储藏，尽管要花费一定的成本（运转冷冻设备）。因而，浓缩橙汁的价格不仅取决于当前的供给，还依赖于对未来供给的预期。今天的霜冻意味着未来浓缩橙汁的供给会下降，但由于可以使用现有的浓缩橙汁来满足现在或未来的需求，因此现有浓缩橙汁的价格现在就会上涨。但是，当发生霜冻时，浓缩橙汁的价格确切地上涨多少？这个问题的答案不仅橙汁交易者感兴趣，更一般地说，研究现代商品市场运行的经济学家们也感兴趣。要弄清楚橙汁的价格是如何对天气的变化做出反应的，我们必须分析橙汁价格和天气的数据。

从1950 年1 月到2000 年12 月的冷冻浓缩橙汁价格的月度数据，及其月度百分比变化和佛罗里达州橙子种植区的气温变化，绘制在图13—1 中。图13—1(a)中所绘制的价格是批发商所支付的冷冻浓缩橙汁的实际平均价格。这个价格用产成品的整体生产者价格指数进行了缩减，以剔除整体价格水平通货膨胀的影响。绘制在图13.1(b)中的价格百分比变化是橙汁在月度间价格的变化。绘制在图13—1(c)中的气温数据是在佛罗里达州的奥兰多机场测量的“冷冻温度日”(freezing degree days)数，用在某月份给定日期里华氏最低气温低于冰点的度数之和来计算。例如，在1950 年11 月，机场气温两次降到冰点以下，在25 日(31°)和29 日(29°)共计4 个冷冻温度日((32 - 31) + (32 - 29))（该数据在附录13.1 中被更为详细地描述）。通过比较图13—1 中的图形可以看到，浓缩橙汁的价格波动很大，其中一部分波动看起来与佛罗里达州的寒冷天气有关。

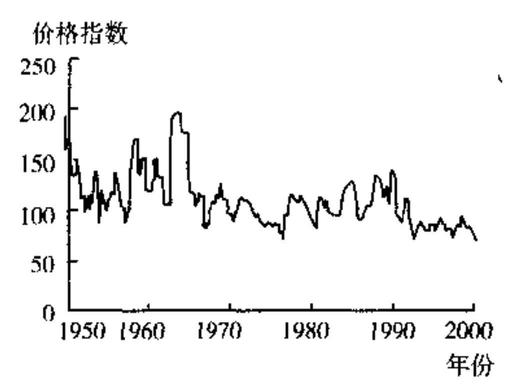
我们现在开始分析橙汁价格和天气之间的定量关系，使用天气变冷时对橙汁价格上涨量有影响的回归模型。因变量是该月期间橙汁价格的百分比变化（$\%ChgP_t$，这里$\%ChgP_t = 100 \times \Delta\ln(p_t^{OJ})$，$p_t^{OJ}$ 是橙汁的实际价格）。回归因子是那个月份里的冷冻温度日数(FDD_t)。使用1950 年1 月到2000 年12 月的月度数据估计这个回归方程（本章中的所有回归都用一样的数据），共计$T = 612$ 个观测值。

$$\widehat{\%ChgP_t} = \underset{(0.22)}{-0.40} + \underset{(0.13)}{0.47}\ FDD_t \qquad (13.1)$$

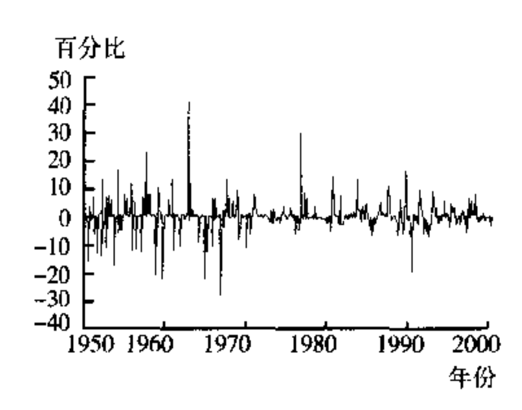
本节所报告的标准误不是通常的OLS 标准误，而是适合于当误差项和回归因子都自相关时的异方差—自相关——致性(HAC)标准误。HAC 标准误在13.4 节中讨论，这里只是使用，没有给出进一步的解释。

根据这个回归，在某个月份，多一个冷冻温度日会使该月份浓缩橙汁的价格上升

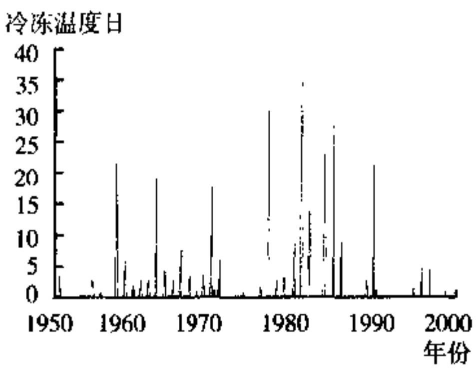
第4部分



(a) 冷冻浓缩橙汁的价格指数



(b) 冷冻浓缩橙汁价格的百分比变化



(c) 佛罗里达州奥兰多的月度冷冻温度日

注:冷冻浓缩橙汁的价格在各月之间一直存在很大的波动。许多大的波动与橙子种植园的故乡奥兰多的霜冻天气一致。

图 13—1 橙汁的价格与佛罗里达州的天气(1950—2000 年)

0.47%。在有 4 个冷冻温度日的月份里,如 1950 年 11 月,相对于没有低于冰点日子的月份而言,估计浓缩橙汁的价格会增加 1.88%(4×0.47%)。

由于公式(13.1)中的回归只包括对同期天气的测度,因此,它没有捕捉到寒潮对下一月份橙汁价格任何延续的影响。为了捕捉这些延续的影响,我们需要考虑 *FDD* 的同期值及其滞后值对价格的影响,这可以通过扩展公式(13.1)中的回归来实现,例如,用 *FDD* 在前 6 个月里的滞后值。

$$\begin{aligned}\widehat{\%ChgP_t} = &\underset{(0.23)}{-0.65} + \underset{(0.14)}{0.47}\,FDD_t + \underset{(0.08)}{0.14}\,FDD_{t-1} + \underset{(0.06)}{0.06}\,FDD_{t-2} + \underset{(0.05)}{0.07}\,FDD_{t-3} \\ &+ \underset{(0.03)}{0.03}\,FDD_{t-4} + \underset{(0.03)}{0.05}\,FDD_{t-5} + \underset{(0.04)}{0.05}\,FDD_{t-6}\end{aligned} \tag{13.2}$$

公式(13.2)就是一个分布滞后回归。公式(13.2)中 FDD_t 的系数估计了在霜冻发生的月份里价格的百分比增加,多一个冷冻温度日估计会使该月份的价格增加 0.47%。FDD_t 一阶滞后 FDD_{t-1} 的系数估计了由上个月的冷冻温度日所引起的价格的百分比增加。二阶滞后系数估计了两个月以前冷冻温度日的影响,依此类推。或者说,*FDD* 的一阶滞后系数估计了在霜冻发生后的一个月里 *FDD* 增加 1 单位的影响。因此,公式(13.2)中的估计系数是 FDD_t 增加 1 单位对 $\%ChgP_t$ 的当前值和未来值产生影响的估计值,也就是说,它们是 FDD_t 对 $\%ChgP_t$ 动态因果效应的估计值。例如,据估计,1950 年 11 月的 4 个冷冻温度日已使 1950 年 11 月份期间橙汁的价格增加了 1.88%,在 1950 年 12 月再额外增加 0.56%(4×0.14%),在 1951 年 1 月再额外增加 0.24%(4×0.06%),等等。



13.2 动态因果效应

在学习更多的估计动态因果效应的工具之前，我们应该花一些时间来考虑动态因果效应的精确含义是什么。清楚地理解动态因果效应的含义有助于更清楚地理解它能够被估计出来的条件。

13.2.1 因果效应和时间序列数据

1.2 节将因果效应定义为一个理想的随机化控制实验的结果。当一位园艺家随机地向一些西红柿地块而不是其他的西红柿地块施肥，然后测量产量，在施肥的和未施肥的地块之间产量的预期差就是施肥对西红柿产量的效应。不过，这个实验概念是一个有多个主体（多个西红柿地块或多个人）的概念，因此数据要么是截面的（在收获结束时西红柿的产量），要么是面板数据（在一个实验性的工作培训项目前后个人的收入）。由于有多个主体，因此既有处理组，也有控制组，进而就可以估计处理的因果效应了。

在时间序列应用中，这个根据理想的随机化控制实验定义的因果效应需要修正，具体来讲，就是考虑宏观经济中的一个重要问题：估计一个不可预测的短期利率的变化对给定国家当前和未来经济行为的影响，经济行为用 GDP 测度。照字面意思理解，根据 1.2 节的随机化控制实验，我们必须将不同的经济总体随机地分配到处理组和控制组。处理组中的中央银行会使用随机利率变化这个处理，而控制组中的中央银行不使用这个随机变化的处理。对这两个组而言，经济活动（例如 GDP）将在接下来的几年里被测量，但如果我们对估计特定国家（比如说，美国）的这种效应感兴趣，那么我们应该怎样做呢？于是，这个实验会要求作为主体的美国有不同的“复本”，并将一些复本分配到处理组，将另一些复本分配到控制组。显然，这个“并行总体”的实验是不可行的。



相反，在时间序列数据中，所考虑的随机化控制实验是由相同的主体（比如说，美国经济）在不同的时点（20 世纪 70 年代、80 年代等等）被给予不同的处理（随机选择的利率的变化）。在这个框架下，不同时期的这个单个主体既扮演处理组的作用，也扮演控制组的作用：有时美联储会改变利率，而有时又不会。因为数据是随时间搜集的，所以使测度动态因果效应（即处理对所研究结果影响的时间路径）成为可能。例如，短期利率意外增加 2 个百分点，且维持了一个季度，它最初对产量的影响可能是微乎其微的；在两个季度以后，GDP 增长可能会减缓，在一到一年半以后减速最大；而在接下来的两个年度里，GDP 增长可能会回到正常水平上。这个因果效应的时间路径就是利率的意外变化对 GDP 增长的动态因果效应。

作为第二个例子，考虑冷冻温度日对橙汁价格的因果效应。可以想象，有许多假设的实验，每个实验得到不同的因果效应，这是完全可能的。其中的第一个实验是，改变佛罗里达州橙子林的天气，但保持其他地方的天气不变，例如，保持德克萨斯州的葡萄园和其他柑橘类水果种植区的天气不变。这个实验在保持其他地区天气不变的情况下测度了一个局部的效应。第二个实验可以改变所有地区的天气，在这里“处理”是整个天气模式的应用。如果天气在有竞争性的作物种植区之间是相关的，那么这两个动态因果效应的结果就会不同。在本章中，我们考虑后一个实验中的因果效应，即应用整个天气模式的因果效应。这个实验和佛罗里达州的因果效应的例子比较接近，即在不保持其他农业区的天气不变的情况下，测量佛罗里达州天气变化对价格的动态效应。



动态效应和分布滞后模型。因为动态效应一定是随时间而发生的，所以，用来估计动态

分布滞后系数构成了非零动态因果效应的全部内容。我们称这个假设——$E(u_t|X_t,X_{t-1},\cdots)=0$——为过去和现在的外生性(past and present exogeneity),但由于该定义与第10章中的外生性定义类似,因此我们只使用术语外生性(exogeneity)。

外生性的第二个概念是,给定X_t的所有过去值、现在值和未来值的条件下,误差项的均值为零,即$E(u_t|\cdots,X_{t+2},X_{t+1},X_t,X_{t-1},X_{t-2},\cdots)=0$,这被称为严外生性(strict exogeneity),为清楚起见,我们还称其为过去、现在和未来的外生性(past, present, and future exogeneity)。引入严外生性概念的原因是,当X是严外生的时,存在比公式(13.3)中的分布滞后模型系数的OLS估计量更有效的动态因果效应系数的估计量。

外生性(过去和现在)与严外生性(过去、现在和未来)之间的差异在于,严外生性在条件期望中包含了X的未来值,因而,严外生性隐含外生性,但反之并不成立。理解这两个概念之间差异的一种方法是,考虑X和u之间相关关系定义的含义。如果X是(过去和现在)外生的,那么u_t与X_t的现在值和过去值均无关。如果X是严外生的,那么除上述无关外,u_t还与X_t的未来值无关。例如,如果Y_t的变化引起X_t的未来值变化,那么X_t不是严外生的,即使它可能是(过去和现在)外生的。

为了具体说明,考虑按公式(13.3)所描述的一个西红柿与肥料之间多年关系的假设实验。因为在该假设实验中肥料是随机地施加的,所以它是外生的。因为现在的西红柿产量不依赖于未来的施肥量,所以施肥时间序列也是严外生的。

作为第二个例子,考虑橙汁价格的例子,其中Y_t是橙汁价格的月度百分比变化,X_t是那个月份中冷冻温度日数。从橙汁市场角度来看,我们可以认为天气——冷冻温度日数——仿佛是被随机分配的,因为天气超出了人力控制的范围。如果FDD的效应是线性的,它在r个月以后对价格没有影响,那么就可以推论天气是外生的。但天气是严外生的吗?给定FDD的未来值,如果u_t的条件均值是非零的,那么FDD就不是严外生的。要回答这个问题,需要仔细考虑u_t中确切地包含了什么。尤其是,如果橙汁市场参与者在决定以某一给定价格买卖多少橙汁时使用FDD的预测值,那么橙汁的价格以及误差项u_t就能够体现未来FDD的信息,它使得u_t成为FDD的一个有用预测因子。这意味着,u_t将与FDD_t的未来值相关。根据这个推理,由于u_t包含了未来佛罗里达州天气的预测,因此FDD会是(过去和现在)外生的,而不是严外生的。这个例子与西红柿和肥料的例子之间的区别是,橙汁市场的参与者受未来佛罗里达州天气预测的影响,而西红柿农作物不受未来施肥的影响。在13.6节中,当我们更详细地分析橙汁价格数据时,我们还会回到FDD是否是严外生的问题上。

外生性的两个定义在重要概念13.1中总结。

重要概念 13.1

分布滞后模型和外生性

在分布滞后模型

$$Y_t=\beta_0+\beta_1X_t+\beta_2X_{t-1}+\beta_3X_{t-2}+\cdots+\beta_{r+1}X_{t-r}+u_t \tag{13.4}$$

中,存在两个不同类型的外生性,即两个不同的外生性条件。

过去和现在外生性(外生性):

$$E(u_t|X_t,X_{t-1},X_{t-2},\cdots)=0 \tag{13.5}$$

过去、现在和未来外生性(严外生性):

$$E(u_t|\cdots,X_{t+2},X_{t+1},X_t,X_{t-1},X_{t-2},\cdots)=0 \tag{13.6}$$

如果X是严外生的,那么它也是外生的,但外生性并不隐含着严外生性。



13.3 含有外生回归因子的动态因果效应的估计

如果 X 是外生的，那么它对 Y 的动态因果效应可以用公式(13.4)中分布滞后模型的 OLS 估计方法进行估计。本节总结了这些 OLS 估计量生成有效统计推断的条件，并且介绍了动态乘数和累积动态乘数。

13.3.1 分布滞后模型的假设

分布滞后模型的四个假设与截面多元回归模型的四个假设类似(见重要概念 5.4)，对时间序列数据来说需要做一些修正。

第一个假设是，X 是外生的，它将截面数据的条件零均值假设扩展到包含 X 所有滞后值的情形。如 13.2 节中所论述的，这个假设意味着公式(13.3)中的 r 个分布滞后系数构成了非零动态因果效应的全部内容。在此意义上，总体回归函数概括了 X 的变化对 Y 的全部动态效应。

第二个假设有两个部分：部分(a)要求这些变量是平稳分布的，部分(b)要求当分离变量的相隔时期数量变大时，这些变量是独立分布的。这个假设与 ADL 模型中的相应假设(重要概念 12.6 中的第二个假设)相同，并且在 12.4 节中对这个假设的讨论也适用于这里。

第三个假设是，变量有八阶以上的非零的和有限的矩。这个假设比本书中其他地方所用的四阶有限矩假设更强。如 13.4 节中所讨论的，这个强假设被用在 HAC 方差估计量背后的数学推导中。

第四个假设和截面数据多元回归模型中的假设相同，即不存在完全多重共线性。

分布滞后回归模型及其假设在重要概念 13.2 中总结。

扩展到额外的 X 变量。分布滞后模型可直接扩展到多个变量 X，只要把额外的变量 X 及其滞后项作为回归因子包含在分布滞后模型中就可以了，而且，为了包含这些额外的回归因子，需要修改重要概念 13.2 中的假设。尽管扩展到多个变量 X 的情形从概念上理解是直接的，但是它的符号表示却很复杂，遮蔽了分布滞后模型估计和推断的主要思想。因此，本章对多个 X 变量的情形没有给予直接的处理，只讨论了一个简单的扩展，即单个变量 X 的分布滞后模型。

13.3.2 自相关的 u_t、标准误和推断

在分布滞后模型中，误差项 u_t 可以是自相关的，也就是说，u_t 可能与它的滞后值相关。在时间序列数据中，出现这个自相关的原因是包含在 u_t 中的那些遗漏因子本身可能是序列相关的。例如，假设对橙汁的需求还依赖于收入，那么影响橙汁价格的一个因素就是收入，具体地讲也就是潜在橙汁消费者的总收入。于是，在橙汁价格变化对冷冻温度日的分布滞后回归中，总收入是一个遗漏变量。不过，总收入是序列相关的：总收入在经济衰退期倾向于下降，而在经济扩张期倾向于上升，因此，收入是序列相关的，而且因为它是误差项的一部分，所以 u_t 也是序列相关的。这个例子很典型，由于 Y 的遗漏的决定性因素本身是序列相关的，因此一般来讲，分布滞后模型中的 u_t 将是序列相关的。

u_t 的自相关性不会影响 OLS 的一致性，也不会引入偏差。不过，如果误差项是自相关的，那么一般而言，通常的 OLS 标准误就是不一致的，必须使用不同的公式。因此，误差项的相关与异方差类似，当误差项确实是异方差的时，仅适用于同方差的标准误就是"错误的"，因为当误差项是异方差的时，使用仅适用于同方差的标准误会导致误导性的统计推



断。同样，当误差项序列相关时，根据独立同分布的误差项所预测的标准误是错误的，因为它们也会导致误导性的统计推断。这个问题的解决方法是，使用异方差—自相关——致性(HAC)标准误，这是13.4节讨论的内容。

重要概念 13.2

分布滞后模型的假设

分布滞后模型在重要概念13.1中给出(公式(13.4))，其中：

1. X 是外生的，即 $E(u_t|X_t, X_{-t}, X_{t-2}, \cdots)=0$。

2. (a)随机变量 Y_t 和 X_t 有个平稳的分布；(b)随着 j 的增大，(Y_t, X_t) 与 (Y_{t-j}, X_{t-j}) 逐渐变成独立的。

3. Y_t 和 X_t 有八阶以上的非零的和有限的矩。

4. 不存在完全多重共线性。

13.3.3 动态乘数和累积动态乘数

动态因果效应的另一个名称是动态乘数。累积动态乘数是直到某一给定滞后期的累积因果效应，因此，累积动态乘数测量 X 的变化对 Y 的累积效应。

动态乘数。X 的单位变化在 h 期以后对 Y 的影响，也就是公式(13.4)中的 β_{h+1}，被称为 h 期动态乘数(dynamic multiplier)。因此，联系 X 和 Y 的动态乘数就是公式(13.4)中 X_t 及其滞后项的系数。例如，β_2 是1期动态乘数，β_3 是2期动态乘数，依此类推。在这个术语中，0期(或同期)动态乘数，或叫冲击效应(impact effect)就是 β_1，它是 X 在同期的变化对 Y 的影响。

由于动态乘数是用OLS回归系数估计的，因此，它们的标准误就是OLS回归系数的HAC标准误。

累积动态乘数。h 期累积动态乘数(cumulative dynamic multiplier)，是在接下来的 h 期里 X 的单位变化对 Y 的累积效应。因而，累积动态乘数是动态乘数的累积和。根据公式(13.4)中分布滞后模型的系数，0期累积乘数是 β_1，1期累积乘数是 $\beta_1+\beta_2$，h 期的累积动态乘数是 $\beta_1+\beta_2+\cdots+\beta_{h+1}$。所有单个动态乘数的和 $\beta_1+\beta_2+\cdots+\beta_{r+1}$ 是 X 的变化对 Y 的累积长期效应，称为长期累积动态乘数(long-run cumulative dynamic multiplier)。

例如，考虑公式(13.2)中的回归。增加一个冷冻温度日的瞬时效应，是浓缩橙汁的价格上涨0.47%。一个价格变化在下个月里的累积效应，等于冲击效应与未来一个月动态效应的和，因此，价格的累积效应是初始增量0.47%加上随后较小的增量0.14%，共计0.61%。同样，未来两个月的累积动态乘数等于0.47%+0.14%+0.06%=0.67%。

累积动态乘数可以直接用公式(13.4)中分布滞后模型的修正形式进行估计，这个修正的回归是：

$$Y_t=\delta_0+\delta_1\Delta X_t+\delta_2\Delta X_{t-1}+\delta_3\Delta X_{t-2}+\cdots+\delta_r\Delta X_{t-r+1}+\delta_{r+1}\Delta X_{t-r}+u_t \quad (13.7)$$

公式(13.7)中的系数 $\delta_1, \delta_2, \cdots, \delta_{r+1}$ 实际上就是累积动态乘数。这可以用一点代数知识来证明(见练习13.5)，它证明了公式(13.7)和公式(13.4)中的总体回归是等价的，其中 $\delta_0=\beta_0$，$\delta_1=\beta_1$，$\delta_2=\beta_1+\beta_2$，$\delta_3=\beta_1+\beta_2+\beta_3$，依此类推。$X_{t-r}$ 的系数 δ_{r+1} 是长期动态累积乘数，即 $\delta_{r+1}=\beta_1+\beta_2+\beta_3+\cdots+\beta_{r+1}$。此外，公式(13.7)中系数的OLS估计量与公式(13.4)中OLS估计量对应的累积和相同。例如，$\hat{\delta}_2=\hat{\beta}_1+\hat{\beta}_2$。用公式(13.7)中的设定估计累积动态乘数的主要好处在于，因为回归系数的OLS估计量是累积动态乘数的估计量，所以，公式

(13.7)中系数的 HAC 标准误是累积动态乘数的 HAC 标准误。

13.4 异方差—自相关——致性标准误

如果误差项 u_t 是自相关的,那么 OLS 是一致的,但一般来讲,通常截面数据 OLS 的标准误并不是一致的。这意味着,以通常的 OLS 标准误为基础的常规的统计推断——假设检验和置信区间——一般来讲会产生误导。例如,由 OLS 估计量 ±1.96 倍常规的标准误所构造的置信区间在 95% 的重复样本中不一定会包含真实值,即使样本规模很大。本节首先推导含有自相关误差的 OLS 估计量方差的正确表达式,然后转向异方差—自相关——致性标准误的讨论。

13.4.1 误差项含有自相关的 OLS 估计量的分布

为了使问题简单化,在一个无滞后项的分布滞后回归模型中,考虑 OLS 估计量 $\hat{\beta}_1$,即含有单个回归因子 X_t 的一个线性回归模型:

$$Y_t=\beta_0+\beta_1X_t+u_t \tag{13.8}$$

这里满足重要概念 13.2 的假设。本节证明 $\hat{\beta}_1$ 的方差可被表示成两项之积:在 u_t 不是序列相关的情况下适用的 $\text{var}(\hat{\beta}_1)$ 的表达式,乘以一个修正因子。这个修正因子是因 u_t 的自相关,或更精确地讲,是因 $(X_t-\mu_t)u_t$ 中的自相关而产生的。

如附录 4.3 中所证明的,重要概念 4.2 中 OLS 估计量 $\hat{\beta}_1$ 的公式可被重写为:



$$\hat{\beta}_1=\beta_1+\frac{\frac{1}{T}\sum_{t=1}^{T}(X_t-\bar{X})u_t}{\frac{1}{T}\sum_{t=1}^{T}(X_t-\bar{X})^2} \tag{13.9}$$

其中,公式(13.9)就是公式(4.51),符号有所改变,原来的 i 和 n 被 t 和 T 所代替。由于 $\bar{X}\xrightarrow{p}\mu_X$,且 $\frac{1}{T}\sum_{t=1}^{T}(X_t-\bar{X})^2\xrightarrow{p}\sigma_X^2$,因此,在大样本条件下,$\hat{\beta}_1-\beta_1$ 可近似地由下式表示:

$$\hat{\beta}_1-\beta_1\approx\frac{\frac{1}{T}\sum_{t=1}^{T}(X_t-\mu_X)u_t}{\sigma_X^2}=\frac{\frac{1}{T}\sum_{t=1}^{T}v_t}{\sigma_X^2}=\frac{\bar{v}}{\sigma_X^2} \tag{13.10}$$

其中,$v_t=(X_t-\mu_X)u_t$,且 $\bar{v}=\frac{1}{T}\sum_{t=1}^{T}v_t$,因此:

$$\text{var}(\hat{\beta}_1)=\text{var}\left(\frac{\bar{v}}{\sigma_X^2}\right)=\frac{\text{var}(\bar{v})}{(\sigma_X^2)^2} \tag{13.11}$$

如果 v_t 是独立同分布的,如在重要概念 4.3 中对截面数据所做的假设,那么 $\text{var}(\bar{v})=\text{var}(v_t)/T$,重要概念 4.4 中 $\hat{\beta}_1$ 的方差表达式就是适用的。不过,如果随着时间的推移,u_t 和 X_t 不是独立分布的,那么一般来讲,v_t 将会是序列相关的,因此,重要概念 4.4 中 $\bar{v}$ 方差的表达式就不适用了。相反,如果 v_t 是序列相关的,那么 $\bar{v}$ 的方差可以表示如下:

$$\begin{aligned}\text{var}(\bar{v})&=\text{var}[(v_1+v_2+\cdots+v_T)/T]\\&=[\text{var}(v_1)+\text{cov}(v_1,v_2)+\cdots+\text{cov}(v_1,v_T)+\text{cov}(v_2,v_1)+\text{var}(v_2)+\cdots+\text{var}(v_T)]/T^2\\&=[T\text{var}(v_t)+2(T-1)\text{cov}(v_t,v_{t-1})+2(T-2)\text{cov}(v_t,v_{t-2})+\cdots+2\text{cov}(v_t,v_{t-T+1})]/T^2\\&=\frac{\sigma_v^2}{T}f_T\end{aligned} \tag{13.12}$$

其中:



$$f_T = 1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\rho_j \tag{13.13}$$

这里$\rho_j = \text{corr}(v_t, v_{t-j})$，在大样本条件下，$f_T$趋向于极限，$f_T \to f_\infty = 1 + 2\sum_{j=1}^{\infty}\rho_j$。

当v_t是自相关的时，将公式(13.10)中$\hat{\beta}_1$的表达式和公式(13.12)中var(v)的表达式合并起来，得到$\hat{\beta}_1$的方差表达式。

$$\text{var}(\hat{\beta}_1) = \left[\frac{1}{T}\frac{\sigma_v^2}{(\sigma_X^2)^2}\right]f_T \tag{13.14}$$

其中，f_T在公式(13.13)中给出。

公式(13.14)将$\hat{\beta}_1$的方差表示为两项之积。方括号中第一项是重要概念4.4中给出的$\hat{\beta}_1$的方差表达式，它适用于不存在序列相关的情形。第二项是因子f_T，它是调整序列相关的一个表达式。由于公式(13.14)中这个额外的因子f_T，如果误差项是序列相关的，那么用重要概念4.4中的公式计算的OLS标准误就不正确，更确切地讲，如果$v_t = (X_t - \mu_X)u_t$是序列相关的，那么由于因子f_T的原因，方差估计量就会是不准确的。

13.4.2 HAC标准误

如果公式(13.13)中定义的因子f_T是已知的，那么$\hat{\beta}_1$的方差可以用通常的截面数据方差估计量乘以f_T来估计。不过，这个因子依赖于v_t这个未知的自相关关系，因此必须估计它。不论是否存在异方差，不论v_t是否是自相关的，加入这种调整的$\hat{\beta}_1$的方差估计量都是一致的。因此，这个估计量被称为$\hat{\beta}_1$方差的异方差—自相关——致性(heteroskedasticity-and autocorrelation-consistent，简写为HAC)估计量，HAC方差估计量的平方根是$\hat{\beta}_1$的HAC标准误(HAC standard error)。



HAC方差表达式。$\hat{\beta}_1$方差的异方差—自相关——致性估计量是：

$$\tilde{\sigma}^2_{\hat{\beta}_1} = \hat{\sigma}^2_{\hat{\beta}_1}\hat{f}_T \tag{13.15}$$

其中，$\hat{\sigma}^2_{\hat{\beta}_1}$是不存在序列相关时公式(4.19)中给出的$\hat{\beta}_1$的方差估计量，$\hat{f}_T$是公式(13.13)中因子$f_T$的估计量。

构造一致性估计量$\hat{f}_T$的工作是富有挑战性的。为什么？我们考虑两种极端的情况。一种极端情况，根据公式(13.13)中的公式，用样本自相关系数$\hat{\rho}_j$(公式(12.6)中所定义的)代替总体自相关系数ρ_j看上去似乎是很自然的，得到估计量$1 + 2\sum_{j=1}^{T-1}\left(\frac{T-j}{T}\right)\hat{\rho}_j$，但这个估计量包含了许多估计的自相关，所以它是不一致的。直观上讲，由于每个所估计的自相关系数都包含估计误差，因此，通过估计如此多的自相关系数，即使在大样本条件下，这个f_T估计量中的估计误差仍旧会很大。另一种极端情况，人们可以设想只使用几阶样本自相关的情形，比如说只使用一阶样本自相关，而忽略所有的更高阶自相关。虽然这个估计量排除了估计太多自相关系数的麻烦问题，但是它又有了另一个不同的问题：它是不一致的，因为它忽略了出现在公式(13.13)中的额外自相关。简而言之，使用太多的样本自相关会使估计量有很大的方差，但使用太少的自相关会忽略高阶的自相关。因此，在任一种极端情况中，这个估计量都是不一致的。

实际中所使用的f_T的估计量，需要在这两个极端情况之间做出平衡，选择多少阶的自

相关被包括进来，取决于样本规模 T。如果样本规模较小，那么只使用少数自相关，但如果样本规模较大，那么应包括更多的自相关（但仍旧远小于 T）。具体地讲，设 $\hat{f}_T$ 由下式给出：

$$\hat{f}_T = 1 + 2\sum_{j=1}^{m-1}\left(\frac{m-j}{m}\right)\tilde{\rho}_j \tag{13.16}$$

其中，$\tilde{\rho}_j = \sum_{t=j+1}^{T}\hat{v}_t\hat{v}_{t-j} / \sum_{t=1}^{T}\hat{v}_t^2$，$\hat{v}_t = (X_t - \overline{X})\hat{u}_t$（如在 $\hat{\sigma}^2_{\hat{\beta}_1}$ 的定义中那样）。公式(13.16)中的参数 m 被称为 HAC 估计量的截断参数（truncation parameter），因为自相关的和被缩短为或被截断为只包含 $m-1$ 阶自相关，代替出现在公式(13.13)的总体表达式中的 $T-1$ 阶自相关。

要使 $\hat{f}_T$ 是一致的，就必须选择 m，使它在大样本条件下很大，尽管仍比 T 小很多。实际中选择 m 的一个指导原则是使用公式(13.17)，四舍五入到整数。

$$m = 0.75T^{1/3} \tag{13.17}$$

这个公式是建立在这个假定之上的，即 v_t 中有中度的自相关的量，它给出了确定 m（m 作为回归中观测期数的一个函数）的一个基本准则[①]。

由公式(13.17)所得到的截断参数 m 的值，可以根据你所掌握的有关该序列的知识进行修正。一方面，如果 v_t 中存在大量的序列相关，那么就要增大 m 超出公式(13.17)中所给的值。另一方面，如果 v_t 中存在很少量的自相关，那么你可以减小 m。由于 m 的选择存在歧义，因此一个比较好的实用方法是，对一个设定至少尝试一个或两个可供选择的 m 值，确信所得的结果对 m 不敏感为止。

公式(13.15)中的 HAC 估计量中 $\hat{f}_T$ 由公式(13.16)给出，该 HAC 估计量被称为**纽韦—韦斯特方差估计量**（**Newey-West variance estimator**），以提出该估计量的经济计量学家 White Newey 与 Kenneth West 来命名。他们证明了，当与公式(13.17)中那样的规则一起使用时，在通常的假设下，这个估计量是 $\hat{\beta}_1$ 的方差的一个一致估计量（Newey 与 West，1987）。他们的证明（以及 Andrews(1991)中的证明）假设 v_t 有四阶以上的矩，这反而又隐含在 X_t 和 u_t 有八阶以上的矩中，这就是重要概念 13.2 中的第三个假设为 X_t 和 u_t 有八阶以上的矩的原因。

其他 HAC 估计量。纽韦—韦斯特方差估计量并不是惟一的 HAC 估计量。例如，公式(13.16)中的权重 $(m-j)/m$ 可被不同的权重代替。如果使用另一个不同的权重，那么公式(13.17)中选择截断参数的规则就不再适用，应该使用由这些权重所引申出来的不同的规则。使用其他权重对 HAC 估计量的进一步讨论超出了本书的范围，要了解有关这个主题的更多信息，请见 Hayashi(2000，6.6 节)。

扩展到多元回归。本节中所讨论的全部问题可推广到重要概念 13.1 中含有多个滞后项的分布滞后模型，更一般地说，可推广到含有序列相关的误差项的多元回归模型。尤其是，如果误差项是序列相关的，那么通常的 OLS 标准误就是统计推断的一个不可靠的基础，应该用 HAC 标准误代替。如果所用的 HAC 方差估计量是纽韦—韦斯特估计量（基于权重 $(m-j)/m$ 的 HAC 方差估计量），那么不论回归中存在单个回归因子还是多个回归因子，都可以按照公式(13.17)中的规则选择截断参数 m。多元回归中 HAC 标准误的表达式已被嵌入到时间序列数据分析的现代回归软件中。由于这个表达式涉及矩阵代数，因此这里省略了它。数学细节，读者可参考 Hayashi(2000，6.6 节)。

① 如果 u_t 和 X_t 是一阶自相关系数为 0.5 的一阶自相关过程，那么方程(13.17)给出了 m 的"最优"选择，这里的"最优"意味着该估计量可使 $E(\tilde{\sigma}^2_{\hat{\beta}_1} - \sigma^2_{\hat{\beta}_1})^2$ 最小。方程(13.17)是以 Andrews(1991，方程(5.3))推导出的一个更一般的公式为基础的。



HAC 标准误在重要概念 13.3 中总结。

重要概念 13.3

HAC 标准误

问题：在重要概念 13.1 中，分布滞后回归模型中的误差项 u_t 可能是序列相关的。如果是这样的话，那么 OLS 系数估计量是一致的，但一般来讲，通常的 OLS 标准误则不是一致的，它会导致误导性的假设检验和置信区间。

解决方法：标准误应该用该方差的异方差—自相关——致性（HAC）的估计量来计算。HAC 估计量涉及 $m-1$ 个自协方差及方差的估计值。在单个回归因子的情况下，相关表达式在公式（13.15）和公式（13.16）中给出。

在实践中，使用 HAC 标准误需要选择截断参数 m，为此，需使用公式（13.17）中的表达式作为一个基准，然后根据回归因子和误差项有较大或较小的序列相关来增大或减小 m。

13.5 含有严外生回归因子时动态因果效应的估计

当 X_t 是严外生的时，可以有两个可供选择的动态因果效应估计量。第一个这样的估计量，涉及估计一个自回归分布滞后（ADL）模型，而不是分布滞后模型，并依据所估计的 ADL 系数计算动态乘数。这种方法要求估计的系数比分布滞后模型的 OLS 估计的系数少，因而潜在地降低了估计误差。第二种方法是，使用广义最小二乘（generalized least squares，简写为 GLS）法而不是 OLS 法估计分布滞后模型的系数。尽管在分布滞后模型中，用 GLS 法所估计的系数个数和用 OLS 法所估计的系数个数一样，但是 GLS 估计量具有更小的方差。为了使说明简单，先列出这两种估计方法，并在含有单个滞后项和 AR(1) 误差的分布滞后模型的条件下进行讨论。不过，当有许多滞后项出现在分布滞后模型中时，这两个估计量的潜在优势最大，因此，这些估计量就可以推广到含有更高阶自回归误差的广义分布滞后模型。



13.5.1 含 AR(1) 误差项的分布滞后模型

假设 X 的变化对 Y 的因果效应只持续两期，即它含有初始时的冲击效应 β_1 和下一期效应 β_2，但不存在随后的效应，那么合适的分布滞后模型就是只含有 X_{t-1} 的当前值和过去值的那个分布滞后模型。

$$Y_t=\beta_0+\beta_1X_t+\beta_2X_{t-1}+u_t \tag{13.18}$$

如 13.2 节所论述的，一般地说，公式（13.18）中的误差项 u_t 是序列相关的。这个序列相关的一个后果是，如果用 OLS 估计这些分布滞后系数，那么基于通常的 OLS 标准误的统计推断可能是误导性的。正是这个原因，13.3 节和 13.4 节强调了当用 OLS 估计公式（13.18）中的 β_1 和 β_2 时应当使用 HAC 标准误。

在本节，我们采取一种不同的方法来分析 u_t 的序列相关。这种方法（在 X_t 是严外生的情况下是可行的）涉及采用自回归模型分析 u_t 中的序列相关，还涉及在分布滞后模型中使用这个 AR 模型推导出一些可能比 OLS 估计量更有效的估计量。



具体来说，假设 u_t 满足 AR(1) 模型。

$$u_t=\phi_1u_{t-1}+\tilde{u}_t \tag{13.19}$$

其中，ϕ_1 是自回归参数，$\tilde{u}_t$ 是序列无关的，而且这里不需要截距项，因为 $E(u_t)=0$。公式（13.18）和公式（13.19）意味着含有序列相关误差项的这个分布滞后模型可被重写为含有

13.5.3 GLS 估计

当 X_t 是严外生的时，估计动态乘数的第二个策略是使用广义最小二乘(GLS)法，这必须要估计公式(13.23)。为了描述 GLS 估计量，我们一开始假设 ϕ_1 是已知的。因为在实践中它是未知的，这个估计量是不可行的，所以被称为不可行的 GLS 估计量。不过，可用 ϕ_1 的估计量修正这个不可行的 GLS 估计量，得到一个可行的 GLS 估计量。

不可行的 GLS。假设 ϕ_1 是已知的，那么准差分变量 $\widetilde{X}_t$ 和 $\widetilde{Y}_t$ 可以直接地计算。正如在公式(13.24)和公式(13.26)的上下文中所讨论的，如果 X_t 是严外生的，那么 $E(\widetilde{u}_t|\widetilde{X}_t,\widetilde{X}_{t-1},\cdots)=0$。因此，如果 X_t 是严外生的，且 ϕ_1 是已知的，那么公式(13.23)中的系数 α_0,β_1 和 β_2 就可用 $\widetilde{Y}_t$ 对 $\widetilde{X}_t$ 和 $\widetilde{X}_{t-1}$ 的 OLS 回归(包含截距项)来估计，相应所得的 β_1 和 β_2 的估计量——即在 ϕ_1 已知时，公式(13.23)中斜率系数的 OLS 估计量——是不可行的 GLS 估计量(infeasible GLS estimator)。这个估计量之所以不可行，是因为 ϕ_1 是未知的，因此，$\widetilde{X}_t$ 和 $\widetilde{Y}_t$ 计算不出来，进而实际中这些 OLS 估计量也不能计算。

可行的 GLS。可行的 GLS 估计量(feasible GLS estimator)是通过用 ϕ_1 的一个初始估计量 $\hat{\phi}_1$，以计算所估计的准差分，来修正不可行的 GLS 估计量得到的。具体地讲，β_1 和 β_2 的可行的 GLS 估计量就是公式(13.23)中系数 β_1 和 β_2 的 OLS 估计量，用 $\hat{\widetilde{Y}}_t$ 对 $\hat{\widetilde{X}}_t$ 和 $\hat{\widetilde{X}}_{t-1}$ 的回归(含截距项)来计算，这里 $\hat{\widetilde{X}}_t=X_t-\hat{\phi}_1X_{t-1}$，$\hat{\widetilde{Y}}_t=Y_t-\hat{\phi}_1Y_{t-1}$。

初始估计量 $\hat{\phi}_1$ 可以这样来计算：首先用 OLS 估计公式(13.18)中的分布滞后模型，然后使用 OLS 估计公式(13.19)中的 ϕ_1，其中以 OLS 残差 $\hat{u}_t$ 代替不可观测的回归误差 u_t。该 GLS 估计量的这种形式，被称为柯赫伦—奥克特(Cochrane - Orcutt)(1949)估计量。

柯赫伦—奥克特方法的一个扩展方法就是不断重复这个过程：用 β_1 和 β_2 的 GLS 估计量计算 u_t 的修正的估计量；用这些新的残差重新估计 ϕ_1；用这个修正的 ϕ_1 估计量计算修正的所估计的准差分；用这些修正的所估计的准差分重新估计 β_1 和 β_2；继续这个过程，直到 β_1 和 β_2 的估计量收敛为止。这个估计量被称为迭代的柯赫伦—奥克特估计量。

GLS 估计量的非线性最小二乘解释。对 GLS 估计量的一个等价的解释是，在对公式(13.22)强加了参数约束的情况下，它估计了公式(13.21)中的 ADL 模型。这些约束是初始参数 β_0,β_1,β_2 和 ϕ_1 的非线性函数，所以不能用 OLS 执行这个估计，相反，可以使用非线性最小二乘法(NLLS)来估计这些参数。如 9.3 节中所论述的，NLLS 可使所估计的由回归函数所造成的误差平方和达到最小，进而可以认识到该回归函数是所估计的参数的一个非线性函数。一般来说，为了最小化未知参数的非线性函数，NLLS 估计可能需要复杂的算法。不过在我们目前特殊的例子中，不需要那些复杂的算法，而是可以用上文所描述的迭代的柯赫伦—奥克特估计量的算法计算 NLLS 估计量。因此，这个迭代的柯赫伦—奥克特估计量实际上是在公式(13.22)中的非线性约束条件下 ADL 系数的 NLLS 估计量。

GLS 有效性。GLS 估计量的优点在于，当 X 是严外生的且变换后的误差 $\widetilde{u}_t$ 是同方差的时，它在线性估计量中是有效的估计量，至少在大样本条件下。要理解这一点，首先考虑不可行的 GLS 估计量。如果 $\widetilde{u}_t$ 是同方差的，ϕ_1 是已知的(因此 $\widetilde{X}_t$ 和 $\widetilde{Y}_t$ 可以像它们被观测到的那样来处理)，X_t 是严外生的，那么高斯—马尔可夫定理隐含着，公式(13.23)中 α_0,β_1 和

β_2 的 OLS 估计量在全部线性条件无偏估计量中是有效的，也就是说，公式(13.23)中系数的 OLS 估计量是最优线性无偏估计量或 BLUE(见 4.9 节)。由于公式(13.23)中的 OLS 估计量是不可行的 GLS 估计量，因此，这意味着该不可行的 GLS 估计量是 BLUE 的。可行的那个 GLS 估计量与不可行的 GLS 估计量类似，不同的是 ϕ_1 被估计了。由于 ϕ_1 的估计量是一致的，它的方差与 T 成反比，所以在大样本条件下这些可行的和不可行的 GLS 估计量具有相同的方差。在这个意义上，如果 X 是严外生的，那么这个可行的 GLS 估计量在大样本条件下就是 BLUE 的。尤其是，如果 X 是严外生的，那么 GLS 比 13.3 节中所论述的那个分布滞后模型的 OLS 估计量更有效。

这里所给出的柯赫伦—奥克特估计量和迭代的柯赫伦—奥克特估计量是 GLS 估计的特例。一般地说，GLS 估计涉及变换回归模型，使得误差项是同方差的和序列无关的，然后用 OLS 估计变换后的模型的系数。如果 X 是严外生的，那么其 GLS 估计量就是一致的，并且是 BLUE 的，但如果 X 只是(过去和现在)外生的，这个结论则不成立。GLS 的数学推导涉及矩阵代数，所以将它们延迟到 16.6 节。

13.5.4 含有额外滞后项和 AR(p)误差项的分布滞后模型

前面对公式(13.18)和公式(13.19)中含有 X_t 的单个滞后项和 AR(1)的误差项的分布滞后模型的讨论，也适用于含有多个滞后项和 AR(p)误差项的广义分布滞后模型的情形。

含有自回归误差项的广义分布滞后模型。含有 r 阶滞后和 AR(p)误差项的广义分布滞后模型为：

$$Y_t = \beta_0 + \beta_1 X_t + \beta_2 X_{t-1} + \cdots + \beta_{r+1} X_{t-r} + u_t \tag{13.30}$$

$$u_t = \phi_1 u_{t-1} + \phi_2 u_{t-2} + \cdots + \phi_p u_{t-p} + \tilde{u}_t \tag{13.31}$$

其中，$\beta_1, \cdots, \beta_{r+1}$ 是动态乘数，而 $\phi_1, \cdots, \phi_p$ 是误差项的自回归系数。对于误差项的 AR(p)模型，$\tilde{u}_t$ 是序列无关的。

导出公式(13.21)中 ADL 模型的代数运算表明，公式(13.30)和公式(13.31)隐含着可用 ADL 形式将 Y_t 写为：

$$Y_t = \alpha_0 + \phi_1 Y_{t-1} + \cdots + \phi_p Y_{t-p} + \delta_0 X_t + \delta_1 X_{t-1} + \cdots + \delta_q X_{t-q} + u_t \tag{13.32}$$

其中，$q = r + p$，而 $\delta_0, \cdots, \delta_q$ 是公式(13.30)和公式(13.31)中系数 β 和 ϕ 的函数。公式(13.30)和公式(13.31)中的模型可用准差分形式等价地写成：

$$\tilde{Y}_t = \alpha_0 + \beta_1 \tilde{X}_t + \beta_2 \tilde{X}_{t-1} + \cdots + \beta_{r+1} \tilde{X}_{t-r} + \tilde{u}_t \tag{13.33}$$

其中，$\tilde{Y}_t = Y_t - \phi_1 Y_{t-1} - \cdots - \phi_p Y_{t-p}$，$\tilde{X}_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$。

ADL 系数估计的条件。前面在 AR(1)情况下对 ADL 系数一致性估计条件的讨论，可推广到含有 AR(p)误差的一般模型中。公式(13.33)中的条件零均值假设是指：

$$E(\tilde{u}_t | \tilde{X}_t, \tilde{X}_{t-1}, \cdots) = 0 \tag{13.34}$$

由于 $\tilde{u}_t = u_t - \phi_1 u_{t-1} - \phi_2 u_{t-2} - \cdots - \phi_p u_{t-p}$，$\tilde{X}_t = X_t - \phi_1 X_{t-1} - \cdots - \phi_p X_{t-p}$，这个条件等价于：

$$E(u_t | X_t, X_{t-1}, \cdots) - \phi_1 E(u_{t-1} | X_t, X_{t-1}, \cdots) - \cdots - \phi_p E(u_{t-p} | X_t, X_{t-1}, \cdots) = 0 \tag{13.35}$$

为使公式(13.35)对全部的 $\phi_1, \cdots, \phi_p$ 值都成立，必有公式(13.35)中的每一个条件期望都为零，或者说，必有下述条件成立：

$$E(u_t | X_{t+p}, X_{t+p-1}, X_{t+p-2}, \cdots) = 0 \tag{13.36}$$

这个条件不能被 X_t 是(过去和现在)外生的这个条件所隐含，但却可以被 X_t 是严外生

的这个条件所隐含。实际上,在极限意义下,当 p 无穷大时(因此分布滞后模型中的误差项满足无穷阶的自回归),公式(13.36)中的条件变为重要概念 13.1 中的严外生性的条件。

用 OLS 估计 ADL 模型。正像在含有单个滞后项和 AR(1)误差项的分布滞后模型中一样,可以根据公式(13.32)中 ADL 系数的 OLS 估计量估计动态乘数。其一般的表达式与公式(13.29)中的表达式类似,但更复杂一些,最好用滞后乘数符号表示,这些表达式在附录 13.2 中给出。实际上,现代的时间序列回归分析软件会自动地为你做这些计算。

用 GLS 来估计。另一方面,动态乘数可以用(可行的)GLS 来估计。这需要对公式(13.33)中准差分设定的系数的 OLS 估计使用所估计的准差分,和 AR(1)中的情况一样,可使用自回归系数 $\phi_1,\cdots,\phi_p$ 的初始估计量计算所估计的准差分。根据前面对 AR(1)情况的讨论,我们说该 GLS 估计量是渐近 BIUE 的。

在严外生条件下,动态乘数的估计在重要概念 13.4 中总结。

使用哪一个估计:OLS 还是 GLS? 对 ADL 系数的 OLS 估计和分布滞后系数的 GLS 估计,这两种估计选择各有优点,也各有缺点。

与分布滞后模型的 OLS 估计相比,ADL 方法的优点在于它能够减少估计动态乘数所需要的参数个数。例如,公式(13.27)中所估计的 ADL 模型导出公式(13.29)中那个无限长的所估计的分布滞后代表式。只含有 r 阶滞后的分布滞后模型实际上是一个更长阶滞后的分布滞后模型的一个近似,从这个意义上说,ADL 模型可以提供一种只用少数未知参数去估计那些阶数很长的滞后项的简单方法。因此在实际中,用比估计公式(13.37)中分布滞后系数的 OLS 估计所需要的 r 值小得多的 p 值和 q 值,去估计公式(13.39)中的 ADL 模型,这是可能的。换句话说,ADL 设定能够对一个长且复杂的分布滞后模型提供一个简洁的或节俭的概括(进一步的讨论请见附录 13.2)。



重要概念 13.4

严外生条件下动态乘数的估计

含有 r 阶滞后和 AR(p)误差项的广义分布滞后模型为:

$$Y_t=\beta_0+\beta_1X_t+\beta_2X_{t-1}+\cdots+\beta_{r+1}X_{t-r}+u_t \tag{13.37}$$

$$u_t=\phi_1u_{t-1}+\phi_2u_{t-2}+\cdots+\phi_pu_{t-p}+\tilde{u}_t \tag{13.38}$$

如果 X_t 是严外生的,那么可以通过首先使用 OLS 方法估计 ADL 模型

$$Y_t=\alpha_0+\phi_1Y_{t-1}+\cdots+\phi_pY_{t-p}+\delta_0X_t+\delta_1X_{t-1}+\cdots+\delta_qX_{t-q}+\tilde{u}_t \tag{13.39}$$

的系数,进而估计动态乘数 $\beta_1,\cdots,\beta_{r+1}$,这里,$q=r+p$;然后使用回归软件计算动态乘数。或者,可用 GLS 估计公式(13.37)中的分布滞后系数来估计动态乘数。

GLS 估计量的优点是,对于在分布滞后模型中给定的一个滞后长度 r 而言,这些分布滞后系数的 GLS 估计量比 OLS 估计量更有效,至少在大样本条件下是如此。不过在实际中,由于 ADL 模型设定允许比 GLS 估计的参数少,所以 ADL 方法还是有优点的。

13.6 橙汁价格与寒冷天气

本节用时间序列回归的工具,对佛罗里达州的气温和橙汁价格的数据做进一步深入的分析。第一,一次霜冻对橙汁价格的影响会持续多久? 第二,在数据所跨越的 51 年里,这个动态效应一直稳定吗? 还是已发生了变化? 如果变化了,它又是如何变化的?

我们首先使用 13.3 节中的方法来估计动态因果效应作为我们分析的开始，即用价格的百分比变化（$\% ChgP_t$）对该月冷冻温度日（FDD_t）及其滞后值回归的这个分布滞后回归系数采用 OLS 进行估计。为了使这个分布滞后估计量是一致的，*FDD* 必须是（过去和现在）外生的。如 13.2 节中所讨论的，这个假设在这里是合理的。人类不能影响天气，因此在实验中我们将天气看做是仿佛被随机分配的应该是合适的。由于 *FDD* 是外生的，我们可用重要概念 13.1 中公式（13.4）的分布滞后模型系数的 OLS 估计来估计这个动态因果效应。

如 13.3 节和 13.4 节中所论述的，分布滞后回归中的误差项可能是序列相关的，因此非常有必要使用 HAC 标准误，因为它对序列相关进行了调整。纽韦—韦斯特标准误的截断参数（13.4 节的符号中的 m）的初始结果，是用公式（13.17）中的规则进行选择的。因为有 612 个月度观测值，根据那个规则，$m=0.75T^{1/3}=0.75\times612^{1/3}=6.37$，而且，因为 m 必须是整数，所以将它向上进到 $m=7$。下面研究标准误对截断参数选择的敏感性。

$\% ChgP_t$ 对 $FDD_t, FDD_{t-1}, \cdots, FDD_{t-18}$ 的分布滞后回归的 OLS 估计结果总结在表 13—1 的第（1）列中。这个回归的系数（表中只报告了一部分）是，在某月中冷冻温度日的数量增加 1 单位对接下来的 18 个月橙汁价格变化（以百分数表示）的动态因果效应的估计值。例如，在冷冻温度日发生的那个月份里，增加一个冷冻温度日估计会使价格增加 0.50%。一个冷冻温度日在随后的月份里对价格的后继效应变小了：在 1 个月以后，这个估计的效应是使价格进一步增加 0.17%；而 2 个月以后，所估计的效应是使价格再增加 0.07%。这个回归的 R^2 为 0.12，这表明橙汁价格中的大部分变差没有被 *FDD* 的当前值和过去值所解释。

表 13—1　　**冷冻温度日（FDD）对橙汁价格变化的动态效应：所估计的部分动态乘数和累积动态乘数**

滞后阶数	(1) 动态乘数	(2) 累积乘数	(3) 累积乘数	(4) 累积乘数
0	0.50 (0.14)	0.50 (0.14)	0.50 (0.14)	0.51 (0.15)
1	0.17 (0.09)	0.67 (0.14)	0.67 (0.13)	0.70 (0.15)
2	0.07 (0.06)	0.74 (0.17)	0.74 (0.16)	0.76 (0.18)
3	0.07 (0.04)	0.81 (0.18)	0.81 (0.18)	0.84 (0.19)
4	0.02 (0.03)	0.84 (0.19)	0.84 (0.19)	0.87 (0.20)
5	0.03 (0.03)	0.87 (0.19)	0.87 (0.19)	0.89 (0.20)
6 ⋮	0.03 (0.05)	0.90 (0.20)	0.90 (0.21)	0.91 (0.21)
12 ⋮	−0.14 (0.08)	0.54 (0.27)	0.54 (0.28)	0.54 (0.28)
18	0.00 (0.02)	0.37 (0.30)	0.37 (0.31)	0.37 (0.30)
月度指示变量	否	否	否	是 $F=1.01$ $(p=0.43)$
HAC 标准误截断参数（m）	7	7	14	7

注：利用（在附录 13.1 中所描述的）1950 年 1 月到 2000 年 12 月的月度数据，共计 $T=612$ 个观测值，用 OLS 法估计所有的回归。因变量是橙汁价格的月度百分比变化（$\% ChgP_t$）。回归（1）是分布滞后回归，其中含有冷冻温度日的月度数量及其 18 阶滞后值，即 $FDD_t, FDD_{t-1}, \cdots, FDD_{t-18}$，所报告的系数是动态乘数的 OLS 估计值。累积乘数是所估计的动态乘数的累积和。所有的回归都包含截距项，这里没有给出。用最后一行给出的截断数计算的纽韦—韦斯特 HAC 标准误在括号中给出。

与表13—1这样的表格相比,动态乘数图可能会更有效地传递信息。表13—1第(1)列中的动态乘数连同它们的95%的置信区间(用所估计的系数±1.96倍HAC标准误计算)一起被绘制在图13—2(a)中。在最初的价格暴涨之后,随后的价格上涨较小,尽管价格在霜冻之后的前6个月中据估计每个月里会有稍微上涨。由图13—2(a)可见,除了第一个月以外,动态乘数在5%的显著性水平下在统计上都不是显著地异于0的,尽管直到第七个月它们还被估计是正的。

表13—1的第(2)列包含了这个设定的累积动态乘数,即第(1)列中所报告的动态乘数的累积和。这些动态乘数连同它们的95%的置信区间一起被绘制在图13—2(b)中。在1个月以后,冷冻温度日的累积效应使价格增加了0.67%;在2个月以后,价格被估计出已上涨0.74%;而在6个月以后,价格被估计出已上涨0.90%。由图13—2(b)可见,这些累积乘数一直增加到第7个月,因为前7个月的单个动态乘数都是正的。在第8个月里,动态乘数是负的,因此橙汁的价格开始从它的峰值缓慢下降。18个月后,价格的累积增加仅为0.37%,也就是说,这个长期的累积动态乘数只有0.37%。这个长期的累积动态乘数在10%的显著性水平下在统计上不是显著地异于0的($t=0.37/0.30=1.23$)。

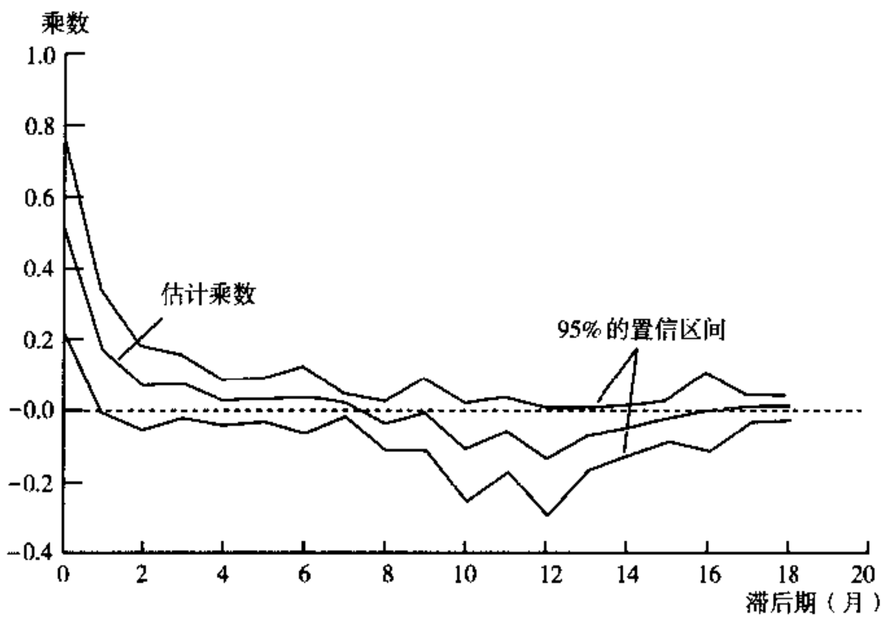
敏感性分析。像在任何实证分析中一样,检查这些结果对该实证分析细节的变化是否敏感,是非常重要的。因此,我们研究这个分析的三个方面:对HAC标准误计算的敏感性;为分析潜在遗漏变量偏差而提供的另一种设定的敏感性;所估计的乘数随时间变化的稳定性分析。

首先,我们研究表13—1的第(2)列中所报告的标准误对HAC截断参数m的不同选择是否敏感。在第(3)列中,报告了$m=14$——是第(2)列中所用值的两倍——的计算结果。该回归设定与第(2)列相同,因此所估计的系数和动态乘数是相等的,只有标准误不同,但碰巧差异不大。我们的结论是:所得结果对HAC截断参数的变化不敏感。

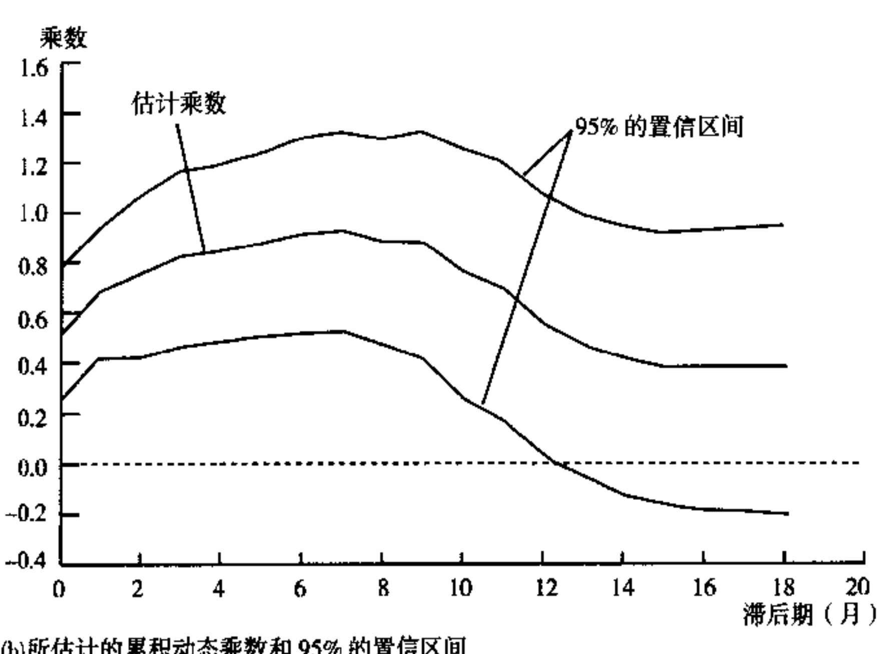
其次,我们研究遗漏变量偏差的一个可能来源。佛罗里达州的霜冻在整个一年里并不是被随机分配的,而是(当然)发生在冬季。如果对橙汁的需求是季节性的(冬季对橙汁的需求是否大于夏季),那么橙汁需求的季节模式可能与*FDD*相关,从而导致遗漏变量偏差。已售出的用于榨橙汁的橙子的数量是内生的:价格和数量由供给和需求的力量同时决定。因而,如7.2节中所论述的,把该数量包含进来将会导致联立偏差。尽管如此,需求中的季节成分可以通过将季节变量作为回归因子来捕捉。因此,表13—1第(4)列中的设定包含了11个月份的二元变量,一个表示该月是否是一月份,一个表示该月是否是二月份,如此等等(和通常一样,必须省略一个二元变量,以防止与截距项的完全多重共线性)。这些月度指示变量在10%的显著性水平下并不是联合地在统计上是显著的($p=0.43$),所估计的累积动态乘数在本质上与不包括月度指标的设定所得出的累积动态乘数相同。总之,需求的季节性波动不是遗漏变量偏差的一个重要来源。

这些动态乘数随时间变化一直稳定吗?① 为了评价动态乘数的稳定性,我们需要检查分布滞后回归的系数随时间的变化是否是稳定的。因为我们心中并没有一个具体的突变日期,所以,我们利用Quandt似然比(QLR)统计量(见重要概念12.9)检验回归系数中的不稳定性。对于全部系数都存在交互作用的第(1)列的回归所计算的QLR统计量(具有15%修匀和HAC方差估计量)的值为9.08,自由度$q=20$(FDD_t及其18阶滞后项和截距项的系数)。表12—5中的1%临界值为2.43,因此QLR统计量在1%的显著性水平下拒绝零假

① 这个小节中稳定性的论述引用了12.7节的内容,如果没有学习那些内容,那么可以跳过这一小节。



(a)所估计的动态乘数和 95% 的置信区间



(b)所估计的累积动态乘数和 95% 的置信区间

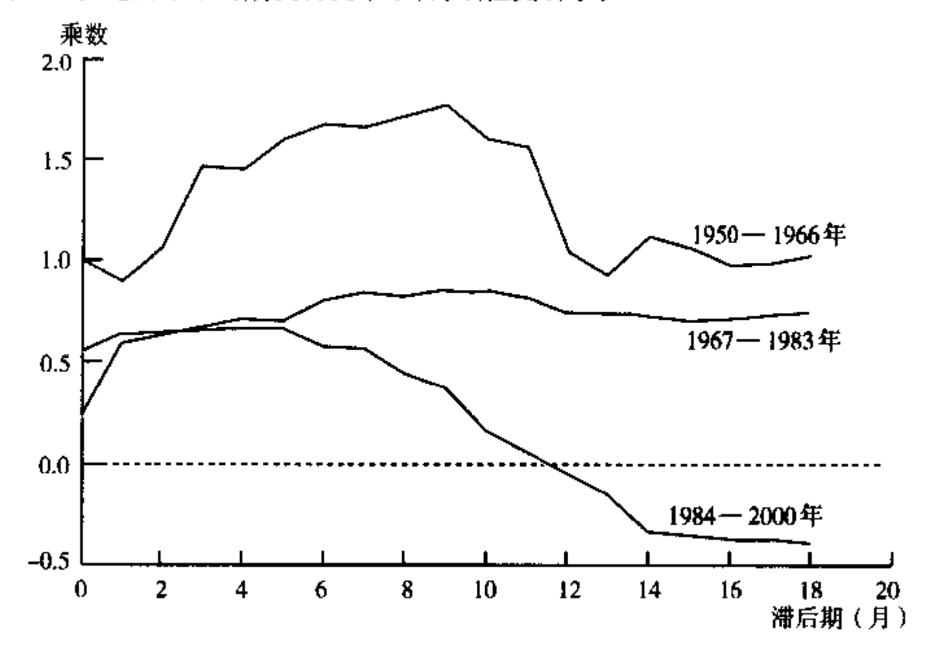
注：所估计的动态乘数表明，霜冻会导致价格的迅速上升。未来价格的上涨比初始冲击小得多。累积动态乘数表明，霜冻对橙汁价格水平具有长期持续性影响，价格在霜冻之后的 7 个月达到峰值。

图 13—2　冷冻温度日（*FDD*）对橙汁价格的动态效应

设。这些 QLR 回归有 40 个回归因子，这是个很大的数；只对六阶滞后重新计算（所以有 16 个回归因子，$q=8$），其结果在 1% 的水平下也拒绝零假设。因此，"动态乘数是平稳的"假设在 1% 的显著性水平下被拒绝。

为了搞清楚动态乘数是怎样随时间的变化而变化的，一种方法是对样本的不同部分计算动态乘数。图 13—3 绘制了对每个子样本进行单独回归所计算的前三分之一（1950—1966）样本、中间的三分之一（1967—1983）样本和最后的三分之一（1984—2000）样本所估计的累积动态乘数。这些估计值显示了一个令人感兴趣并且值得关注的模式。在 20 世纪

50 年代和 60 年代初期，一个冷冻温度日对价格有很大的且持久的影响。一个冷冻温度日对价格变化效应的程度在 20 世纪 70 年代变小了，尽管它仍保持高度的持续性。在 20 世纪 80 年代后期和 90 年代，冷冻温度日的短期效应与 20 世纪 70 年代相同，但持续性变弱了，而且基本上在一年后就消失了。这些估计值表明，佛罗里达州的霜冻对橙汁价格的动态因果效应在 20 世纪后半个时期变得更小了，持续性更弱了。



注：霜冻对橙汁价格的动态效应在 20 世纪后半期发生了显著的变化。在 1950 年到 1966 年期间，霜冻对价格的冲击比后来大，而且在 1984 年到 2000 年期间，霜冻影响的持续性比以前时期弱。

图 13—3　所估计的不同样本期的累积动态乘数



ADL 和 GLS 估计的估计值。如 13.5 节中所论述的，如果分布滞后回归中的误差项是序列相关的，且 *FDD* 是严外生的，那么，比分布滞后系数的 OLS 估计更有效地估计动态乘数是可能的。然而，在使用 GLS 估计量或基于 ADL 模型的估计量之前，我们需要考虑 *FDD* 是否确实是严外生的。的确，人类不能影响天气，但这是否就意味着天气是严外生的呢？给定 *FDD* 的过去值、现在值和未来值，分布滞后回归中的误差项 u_t 是否有条件零均值？

表 13—1 第(1)列中的分布滞后回归所对应的总体回归中的误差项，是价格和以过去 18 个月的天气资料为基础的总体预测值之差。这个差值的产生可有许多原因，其中一个原因是交易者使用了奥兰多的天气预测。例如，如果预测到一个特别寒冷的冬季，那么交易者就会将这个预测综合到价格中，因此，相应的价格就会高出以总体回归为基础的预测值，也就是说，误差项可能是正的。如果这个预测是正确的，那么事实上未来的天气应是寒冷的。因而，未来的冷冻温度日将是正的($X_{t+1}>0$)，而当前的价格会异常地高($u_t>0$)，所以 $\mathrm{corr}(X_{t+1},u_t)$ 是正的。更简单地说，虽然橙汁交易者不能影响天气，但他们能够预测天气(而且确实是如此)（见本章一般兴趣框），结果，价格与天气回归中的误差项同未来的天气是相关的。换句话说，*FDD* 是外生的，但如果这个推理是正确的，那么它就不是严外生的，GLS 和 ADL 估计量也将不会是动态乘数的一致估计量。因此，这些估计量在这个应用中不会被使用。

相反，在一个欧盟向美国的出口对美国收入的回归中，将美国收入看做是外生的观点不太令人信服，因为欧盟居民对美国出口商品的需求构成了对美国出口品总需求的绝大部分。因而，美国对欧盟出口商品需求的下降，会减少欧盟的收入，这反过来又会减少对美国出口商品的需求，进而降低美国的收入。由于这些通过国际贸易建立的联系，欧盟对美国的出口和美国的收入是联立地决定的，因此，在这个回归中可以证明美国的收入不是外生的。这个例子说明了一个更一般的观点，即一个变量是否是外生的取决于它的内容：在一个解释澳大利亚出口的回归模型中，美国的收入看上去是外生的，但在解释欧盟出口的回归模型中却不是如此。

13.7.2 石油价格和通货膨胀

自从20世纪70年代的石油价格上涨以来，宏观经济学家们一直对估计国际原油价格上涨对美国通货膨胀率的动态效应感兴趣。因为石油价格是在大部分石油价格由国外的石油生产国决定这样的世界市场中决定的，所以，起初人们可能会认为石油价格是外生的，但石油价格和天气不一样：OPEC 成员国从战略的角度确定石油的产量，他们考虑到了许多因素，包括世界经济状况。由于石油价格（或产量）是在对当前和未来的世界经济状况（包括美国的通货膨胀）的评估基础上做出的，因此，在这个意义上说，石油价格是内生的。

13.7.3 货币政策和通货膨胀

负责制定货币政策的中央银行家需要知道货币政策对通货膨胀的效应。因为货币政策的主要工具是短期利率（"短利率"），因此，这意味着他们需要知道短期利率的变化对通货膨胀的动态因果效应。尽管短期利率由中央银行决定，但它并不是由中央银行家随机地决定的（在一个理想的随机化实验中就是这样），而是内生地制定的：中央银行根据对当前和未来经济状况的评估，特别是对当前和未来通货膨胀率的评估，制定了短期利率。反过来，通货膨胀率又依赖于利率（高利率降低总需求），但利率依赖于通货膨胀率、它的过去值以及（预期的）未来值。因而，短期利率是内生的，短期利率变化对未来通货膨胀的动态因果效应，不能够用通货膨胀率对当前和过去利率的 OLS 回归一致地估计出来。

13.7.4 菲利浦斯曲线

第12章中研究的菲利浦斯曲线是通货膨胀率的变化对通货膨胀滞后变化和失业率滞后项的回归。由于失业率的滞后项发生在过去，起初人们可能会认为不可能存在当前通货膨胀率对失业率过去值的反馈，因此，可以将失业率的过去值看做是外生的。但失业率的过去值在一个实验中并不是随机分配的，相反，以前的失业率和通货膨胀的过去值是联立地决定的。由于通货膨胀率和失业率是联立地决定的，因此，包含在 u_t 中的决定通货膨胀的其他因素与失业率的过去值相关，也就是说，失业率不是外生的。由此得出，失业率不是严外生的，所以实证上用菲利浦斯曲线（例如，公式（12.17）中的 ADL 模型）所计算的动态乘数不是失业率变化对通货膨胀的动态因果效应的一致性估计值。

13.8 结论

时间序列数据提供了估计 X 的变化对 Y 的效应的这种时间路径的机会，即 X 的变化对 Y 的动态因果效应。然而，要利用一个分布滞后模型估计动态因果效应，X 必须是外生的，

就像它在一个理想的随机化实验中被随机地决定那样。如果 X 不仅是外生的，而且还是严外生的，那么动态因果效应可以用一个自回归分布滞后模型或 GLS 来估计。

在一些应用中，比如估计佛罗里达州霜冻天气对橙汁价格的动态因果效应，可以创造一个令人信服的条件，即回归因子（冷冻温度日）是外生的，因此，该动态因果效应能够用分布滞后系数的 OLS 估计方法来估计。但即使在这个应用中，经济理论指出天气不是严外生的，ADL 或 GLS 方法也是不合适的。此外，在许多经济计量学家感兴趣的关系中存在联立因果关系，所以，这些设定中的回归因子不是外生的，既不是严外生的，也不是其他外生的。要确定回归因子是否是外生的（或严外生的），最终需要结合经济理论、制度知识和专家判断。

总结

1. 时间序列中的动态因果效应是通过一个随机化实验来定义的，在一个随机化实验中，相同的主体（实体）在不同的时间接受不同的随机分配的处理。当 X 的时间路径是随机决定的，并且不受那些能够影响 Y 的其他因素的影响，则 Y 对 X 及其滞后项的分布滞后回归的系数，可以解释为动态因果效应。

2. 在一个 Y 对 X 的当前值和过去值的分布滞后回归中，如果误差 u_t 的条件均值不依赖于 X 的当前值及过去值，那么变量 X 是（过去和现在）外生的。如果除此之外，u_t 的条件均值还不依赖于 X 的未来值，那么 X 是严外生的。



3. 如果 X 是外生的，那么在一个 Y 对 X 的当前值和过去值的分布滞后回归中，回归系数的 OLS 估计量便是动态因果效应的一致性估计量。一般而言，在这个回归中的误差项 u_t 是序列相关的，所以，常规的标准误是误导的，必须使用 HAC 标准误来代替它。

4. 如果 X 是严外生的，那么动态乘数可以用 ADL 模型的 OLS 来估计，也可以用 GLS 来估计。

5. 外生性是一个较强的假设，因为存在联立因果关系，所以它在经济时间序列数据中常常不成立。严外生性的假设是一个更强的假设。

重要术语

动态因果效应　分布滞后模型　外生性　严外生性　动态乘数　冲击效应　累积动态乘数　长期累积动态乘数　异方差—自相关——致性（HAC）标准误　截断参数　纽韦—韦斯特方差估计量　广义最小二乘（GLS）　准差分　不可行的 GLS 估计量　可行的 GLS 估计量

复习概念

13.1　在 20 世纪 70 年代，常见的一个应用是用一个分布滞后模型估计名义国内生产总值（Y）的变化与货币供应量（X）的当前和过去的变化之间的关系。在什么样的假设条件下，这个回归会估计出货币对名义国内生产总值的因果效应？这些假设在像美国这样国家的现代经济中可能满足吗？

13.2　假设 X 是严外生的。一位研究人员估计了 ADL(1,1) 模型，计算了回归残差，并

发现残差是高度序列相关的。该研究人员应该估计一个新的含有额外滞后项的 ADL 模型，还是只对 ADL(1,1)所估计的系数使用 HAC 标准误?

13.3　假设一个分布滞后模型被估计出来了，其中因变量是 ΔY_t 而不是 Y_t。请解释你如何计算 X_t 对 Y_t 的动态乘数。

13.4　假设你在公式(13.2)中添加 FDD_{t+1} 作为一个额外的回归因子，如果 FDD 是严外生的，那么你预期 FDD_{t+1} 的系数应该是 0 还是非 0? 如果 FDD 是外生的而不是严外生的，你的答案会变化吗?

练习

*13.1　在发达国家的几次经济衰退期间，石油价格的上涨一直受到谴责。为了量化石油价格对实际经济活动的影响，研究者们已做了类似本章所讨论的回归。设 GDP_t 表示美国季度国内生产总值，并设 $Y_t=100\ln(GDP_t/GDP_{t-1})$ 为 GDP 的季度百分比变化。James Hamilton，一位经济计量学家和宏观经济学家，他指出只有当石油价格暴涨到它的近期的过去值之上时，它才会反向影响经济。具体来说，设 O_t 等于 0 或日期 t 时的石油价格与其在过去几年中最大值的百分点差异中较大的一个。在 1955：Ⅰ—2000：Ⅳ期间所估计的联系 Y_t 和 O_t 的一个分布滞后模型是：

$$\hat{Y}_t=\underset{(0.1)}{1.0}-\underset{(0.054)}{0.055O_t}-\underset{(0.057)}{0.026O_{t-1}}-\underset{(0.048)}{0.031O_{t-2}}-\underset{(0.042)}{0.109O_{t-3}}-\underset{(0.053)}{0.128O_{t-4}}$$
$$+\underset{(0.025)}{0.008O_{t-5}}+\underset{(0.048)}{0.025O_{t-6}}-\underset{(0.039)}{0.019O_{t-7}}+\underset{(0.042)}{0.067O_{t-8}}$$

a. 假设石油价格在前一次峰值之上暴涨了 25%，并维持在这个新的较高水平上(所以，$O_t=25, O_{t+1}=O_{t+2}=\cdots=0$)，那么在接下来的两年里，对每一季度产出增长的预测效应是什么?

b. 为你在(a)中的答案构造一个 95% 的置信区间。

c. GDP 增长的累积变化在未来 8 个季度里的预测值是多少?

d. 检验 O_t 和它的滞后项的系数是否都为 0 的 HACF 统计量为 3.49。这些系数都显著地异于 0 吗?

13.2　宏观经济学家还注意到，石油价格暴涨之后利率会变化。设 R_t 表示 3 个月短期国债利率(以年利率的百分点表示)，在 1955：Ⅰ—2000：Ⅳ期间所估计的联系 R_t 的变化(ΔR_t)与 O_t 之间关系的分布滞后回归是：

$$\Delta R_t=\underset{(0.06)}{0.07}+\underset{(0.045)}{0.062O_t}+\underset{(0.034)}{0.048O_{t-1}}-\underset{(0.028)}{0.014O_{t-2}}-\underset{(0.169)}{0.086O_{t-3}}-\underset{(0.058)}{0.000O_{t-4}}$$
$$+\underset{(0.065)}{0.023O_{t-5}}-\underset{(0.047)}{0.010O_{t-6}}-\underset{(0.038)}{0.100O_{t-7}}-\underset{(0.025)}{0.014O_{t-8}}$$

a. 假设石油价格在前一次峰值之上暴涨了 25%，并维持在这个新的较高水平上(所以 $O_t=25, O_{t+1}=O_{t+2}=\cdots=0$)，那么在接下来的两年里，每一季度利率的预测变化是什么?

b. 为你在(a)中的答案构造一个 95% 的置信区间。

c. 在时期 $t+8$，石油价格的这种变化对利率水平的效应是多少? 你的答案与累积乘数之间有什么关系?

d. 检验 O_t 和它的滞后项的系数是否都为 0 这一假设的 HACF 统计量为 4.25。这些系数都显著地异于 0 吗?



13.3　考虑两个不同的随机化实验。在实验 A 中，石油价格是随机制定的，中央银行根据它对经济状况（包括石油价格的变化）做出反应的通常政策规则制定政策。在实验 B 中，石油价格是随机制定的，中央银行保持利率不变，尤其是对石油价格的变化不作反应。在这两个实验中，GDP 增长是可观测的。现在假设练习 13.1 的回归中石油的价格是外生的，练习 13.1 中所估计的动态因果效应符合哪一个实验，A 还是 B？

13.4　假设石油价格是严外生的，讨论你如何改善练习 13.1 中动态乘数的估计值。

13.5　根据公式（13.4）推导公式（13.7），证明：$\delta_0=\beta_0,\delta_1=\beta_1,\delta_2=\beta_1+\beta_2,\delta_3=\beta_1+\beta_2+\beta_3$（等等）。（提示：注意 $X_t=\Delta X_t+\Delta X_{t-1}+\cdots+\Delta X_{t-p+1}+\Delta X_{t-p}$）

附录 13.1　橙汁数据集

橙汁价格数据是由美国劳工统计局（BLS 序列 wpu02420301）所搜集的生产者价格指数（PPI）中的加工食品和饲料类中的冷冻橙汁部分。橙汁价格数据序列除以制成品的总体 PPI 来调整总体价格水平的通货膨胀。冷冻温度日序列是用从美国商业部的国家海洋和气象局（NOAA）所得到的奥兰多机场地区的日最低温度记录来构造的。*FDD* 序列的构造使得它的时间安排和橙汁价格数据的时间安排近似地匹配。具体来说，冷冻橙汁的价格数据是在每个月的中旬通过调查一个生产者样本来搜集的，尽管搜集数据的确切日期在各个月份有所不同。因此，*FDD* 序列被构造为从一个月的 11 号到下一个月 10 号的冷冻温度日数，即 *FDD* 是 0 和 32 减去当日最低温度这两个数据中的最大值，然后从当月的 11 号到下月的 10 号对这期间的 *FDD* 求和。因此，2 月份的 $\%ChgP_t$ 就是从 1 月中旬到 2 月中旬实际橙汁价格的百分比变化，而 2 月份的 FDD_t 则是从 1 月 11 日到 2 月 10 日的冷冻温度日数。



附录 13.2　用滞后算子符号表示 ADL 模型与广义最小二乘法

本附录给出了用滞后算子符号表示的分布滞后模型，推导了分布滞后模型的 ADL 和准差分代表式，并讨论了 ADL 模型比初始的分布滞后模型含有更少的参数的条件。

用滞后算子符号表示的分布滞后、ADL 和准差分模型

和附录 12.3 中所定义的一样，滞后算子 L 具有性质 $L^jX_t=X_{t-j}$，并能够将分布滞后 $\beta_1X_t+\beta_2X_{t-1}+\cdots+\beta_{r+1}X_{t-r}$ 表示为 $\beta(L)X_t$，这里 $\beta(L)=\sum_{j=0}^{r}\beta_{j+1}L^j$，其中 $L^0=1$。因此，重要概念 13.1 中的分布滞后模型（公式（13.4））可以用滞后算子符号写为：

$$Y_t=\beta_0+\beta(L)X_t+u_t \tag{13.40}$$

此外，如果误差项 u_t 满足 AR(P)，那么它可被写为：

$$\phi(L)u_t=\tilde{u}_t \tag{13.41}$$

其中，$\phi(L)=\sum_{j=0}^{p}\phi_jL^j,\phi_0=1$，$\tilde{u}_t$ 是序列无关的（注意，这里定义的 $\phi_1,\cdots,\phi_p$ 是公式（13.31）的符号中 $\phi_1,\cdots,\phi_p$ 的负数）。

为了推导 ADL 模型，将公式（13.40）的两边前乘 $\phi(L)$，因此有：

$$\phi(L)Y_t=\phi(L)[\beta_0+\beta(L)X_t+u_t]=\alpha_0+\delta(L)X_t+\tilde{u}_t \tag{13.42}$$



其中：

$$\alpha_0=\phi(1)\beta_0,\delta(\mathrm{L})=\phi(\mathrm{L})\beta(\mathrm{L}),这里的\ \phi(1)=\sum_{j=0}^{p}\phi_j \tag{13.43}$$

要推导准差分模型，注意 $\phi(\mathrm{L})\beta(\mathrm{L})X_t=\beta(\mathrm{L})\phi(\mathrm{L})X_t=\beta(\mathrm{L})\widetilde{X}_t$，这里 $\widetilde{X}_t=\phi(\mathrm{L})X_t$。因此，重新整理公式(13.42)，得到：

$$\widetilde{Y}_t=\alpha_0+\beta(\mathrm{L})\widetilde{X}_t+\widetilde{u}_t \tag{13.44}$$

其中，$\widetilde{Y}_t$ 是 Y_t 的准差分，即 $\widetilde{Y}_t=\phi(\mathrm{L})Y_t$。

ADL 和 GLS 估计量

ADL 系数的 OLS 估计量是用公式(13.42)的 OLS 估计得到的。初始的分布滞后系数是 $\beta(\mathrm{L})$，用所估计的系数来表达，有 $\beta(\mathrm{L})=\delta(\mathrm{L})/\phi(\mathrm{L})$，也就是说，$\beta(\mathrm{L})$ 中的系数满足 $\phi(\mathrm{L})\beta(\mathrm{L})=\delta(\mathrm{L})$ 所隐含的约束条件。因此，基于 ADL 模型系数的 OLS 估计量 $\hat{\delta}(\mathrm{L})$ 和 $\hat{\phi}(\mathrm{L})$ 的动态乘数估计量为：

$$\hat{\beta}^{ADL}(\mathrm{L})=\hat{\delta}(\mathrm{L})/\hat{\phi}(\mathrm{L}) \tag{13.45}$$

教材中公式(13.29)的系数表达式可作为公式(13.45)中的一个特例，在那里 $r=1,p=1$。

可行的 GLS 估计量，是通过获得一个初步的 $\phi(\mathrm{L})$ 估计量计算出估计的准差分，用这些估计的准差分估计公式(13.44)中的 $\beta(\mathrm{L})$，并进行迭代(如果需要的话)直到收敛计算出来的。迭代的 GLS 估计量就是使用公式(13.42)中 ADL 模型的 NLLS 估计所计算出来的 NLLS 估计量，它要受公式(13.43)中参数的非线性约束条件所限制。



正像教材中围绕公式(13.36)的讨论所强调的，要使 X_t(过去和现在)是外生的，只使用上述两种估计方法是不够的，因为只有外生性并不能保证公式(13.36)成立。不过，如果 X 是严外生的，那么公式(13.36)确实成立，并且假定重要概念 12.6 中的假设 2 ~4 成立，则这些估计量是一致的和渐近正态的。此外，通常的(截面的异方差稳健的)OLS 标准误为统计推断提供了一个有效的基础。

用 ADL 模型减少参数。假设分布滞后多项式 $\beta(\mathrm{L})$ 可写成一个滞后多项式之比，即 $\theta_1(\mathrm{L})/\theta_2(\mathrm{L})$，这里 $\theta_1(\mathrm{L})$ 和 $\theta_2(\mathrm{L})$ 都是低阶的滞后多项式，那么公式(13.43)中的 $\phi(\mathrm{L})\beta(\mathrm{L})=\phi(\mathrm{L})\theta_1(\mathrm{L})/\theta_2(\mathrm{L})=[\phi(\mathrm{L})/\theta_2(\mathrm{L})]\theta_1(\mathrm{L})$。如果碰巧 $\phi(\mathrm{L})=\theta_2(\mathrm{L})$，则 $\delta(\mathrm{L})=\phi(\mathrm{L})\beta(\mathrm{L})=\theta_1(\mathrm{L})$。如果 $\theta_1(\mathrm{L})$ 的阶数低，那么 ADL 模型中 X_t 的滞后期数 q 会比 r 小很多。因此，在这些假设下，ADL 模型的估计需要潜在地估计比最初的分布滞后模型少很多的参数。正是在这个意义下，ADL 模型能够获得比分布滞后模型更节省的参数数量(即使用更少的未知参数)。

如这里所推导的，$\phi(\mathrm{L})$ 和 $\theta_2(\mathrm{L})$ 碰巧相同这一假设，在应用中似乎是不可能会发生的一种巧合。不过，ADL 模型能够用较少的几个系数捕捉到动态乘数中大量的形态变化。正是由于这个原因，只要 X 是严外生的，ADL 模型的无约束估计就提供了一种近似等同于一个长分布滞后的(即许多的动态乘数)很有吸引力的方法。

第14章 时间序列回归的其他议题

本章讨论时间序列回归的一些其他话题，我们先从预测开始。第12章考虑了单个变量的预测，然而在实际中，你可能想要预测两个或两个以上的变量，比如通货膨胀率和GDP增长率。14.1节介绍了一个用来预测多元变量的模型，即向量自回归(VAR)，这个模型使用两个或多个变量的滞后值以预测这些变量的未来值。在第12章，我们还集中关注了对未来一个时期(如一个季度)的预测问题，但对未来的2个、3个或更多时期进行预测也是重要的。14.2节中讨论了进行这样预测的方法。

14.3节和14.4节回到12.6节讨论的关于随机性趋势的话题。14.3节介绍了其他一些随机性趋势模型和另一种检验单位自回归根的方法。14.4节介绍了协整的概念，当两个变量拥有共同的随机性趋势时，也就是说，当两个变量都包含随机性趋势，但两个变量的加权差分却不含随机性趋势时就产生了协整。

在一些时间序列数据中，尤其是金融数据中，方差随时间的变化而变化：有时序列会表现出高的波动性，而有时波动性却很低，这样就出现了数据序列中的波动集聚现象。14.5节讨论了波动集聚问题，并引入了处理波动集聚的模型，在这些模型中，预测误差的方差随时间的变化而变化，即预测误差是条件异方差的。条件异方差模型有好几个应用。一个应用是计算预测区间，这里的区间宽度随时间变化而变化，以反映不同时期不确定性的差异。另一个应用是预测一项资产(比如说股票)收益的不确定性，这反过来又可能对评估持有股票的风险有用。



14.1 向量自回归

第12章集中于预测通货膨胀率，但实际上经济预测者也忙于预测其他重要的宏观经济变量，比如说失业率、GDP增长率和利率。一种方法是用12.4节中的方法对每个变量建立单独的预测模型，而另一种方法是设计一个能够预测所有变量的单个模型，它有助于使预

测相互一致。用单个模型预测几个变量的一种方法是使用向量自回归(VAR)。VAR 将单变量自回归推广到多元时间序列变量,即它将单变量自回归推广到时间序列变量的一个“向量”。

14.1.1 VAR 模型

向量自回归(vector autoregression)简称 VAR,它含有两个时间序列变量 Y_t 和 X_t,包括两个方程:在一个方程中因变量是 Y_t,在另一个方程中因变量是 X_t。两个方程中的回归因子是这两个变量的滞后值。更一般地说,含有 k 个时间序列变量的 VAR 由 k 个方程组成,每个方程对应一个变量,所有方程中的回归因子是所有各自变量的滞后值。VAR 方程的系数通过 OLS 方法来估计。

VAR 在重要概念 14.1 中总结。

重要概念 14.1

向量自回归

向量自回归是 k 个时间序列回归的一个集合,其中回归因子是所有 k 个序列的滞后值。一个 VAR 将单变量自回归扩展到时间序列变量的一个列表或时间序列变量的一个“向量”。当每个方程中的滞后期数相同且等于 p 时,该方程系统被称为 VAR(p)。

在有两个时间序列变量 Y_t 和 X_t 的情况下,VAR(p)由两个方程组成:

$$Y_t = \beta_{10} + \beta_{11}Y_{t-1} + \cdots + \beta_{1p}Y_{t-p} + \gamma_{11}X_{t-1} + \cdots + \gamma_{1p}X_{t-p} + u_{1t} \tag{14.1}$$

$$X_t = \beta_{20} + \beta_{21}Y_{t-1} + \cdots + \beta_{2p}Y_{t-p} + \gamma_{21}X_{t-1} + \cdots + \gamma_{2p}X_{t-p} + u_{2t} \tag{14.2}$$

其中,β 和 γ 是未知系数,u_{1t}和 u_{2t}是误差项。

VAR 的假设,就是重要概念 12.6 中的时间序列回归假设,这些假设适用于这里的每个方程。VAR 的系数是用 OLS 对每个方程进行估计的。

VAR 的推断。在 VAR 假设下,OLS 估计量在大样本条件下是一致的,并且具有联合正态分布。因此,统计推断以通常的方式进行,例如,可将系数 95% 的置信区间构造为所估计的系数 ±1.96 倍标准误。

在 VAR 的假设检验中出现了一个新的方面,因为含有 k 个变量的 VAR 是 k 个方程的组合或系统。因而,检验涉及多个方程之间的约束条件的联合假设是可能的。

例如,在公式(14.1)和公式(14.2)的两变量 VAR(p)中,你可能会问正确的滞后长度是 p 还是 $p-1$,也就是说,你可能会问这两个方程中 Y_{t-p}和 X_{t-p}的系数是否都为 0。这些系数都为 0 的零假设是:

$$H_0: \beta_{1p} = 0, \beta_{2p} = 0, \gamma_{1p} = 0, \gamma_{2p} = 0 \tag{14.3}$$

备择假设就是这 4 个系数中至少有一个不为 0。因而,零假设涉及了来自于这两个方程的系数,每个方程有两个系数。

因为在大样本条件下所估计的系数具有联合正态分布,所以,通过计算 F 统计量来检验这些系数的约束条件是可能的。这个统计量的精确表达式是复杂的,因为运算符号必须处理多个方程,所以我们省略了它。实际上,大多数现代软件都包含有检验多个方程系统的系数假设的自动程序。

在 VAR 中应该包含多少个变量? VAR 的每个方程中系数的个数与 VAR 中变量的个数成比例。例如,一个含有 5 个变量和四阶滞后的 VAR 在这 5 个方程的每个方程中将有 21

个系数(5 个变量每个都有四阶滞后,加上一个截距项),共计 105 个系数!估计所有这些系数会增加进入预测中的估计误差的数量,这会导致预测精度的恶化。

一个实际应用中的问题是,要使 VAR 中变量的个数尽量地少,特别是要确保所包含的变量彼此之间是相互关联的,这样它们在相互预测时才是有用的。例如,我们从经验证据和经济理论的组合中知道,通货膨胀率、失业率和短期利率是彼此相关联的,这表明这些变量可以在一个 VAR 模型中互相预测。不过,把一个无关的变量包含在 VAR 中不会增加预测内容,反而会引入估计误差,因此会降低预测精度。

确定 VAR 中的滞后长度。[1] VAR 中的滞后长度既可用 F 检验,也可用信息准则来确定。

一个系统方程的信息准则,扩展了 12.5 节中的单方程信息准则。要定义这个信息准则,我们需要采用矩阵运算符号。设 $\boldsymbol{\Sigma}_u$ 为 VAR 误差项的 $k\times k$ 阶协方差矩阵,并设 $\hat{\boldsymbol{\Sigma}}_u$ 为这个协方差矩阵的估计值,这里 $\hat{\boldsymbol{\Sigma}}_u$ 的 i,j 元素为 $\frac{1}{T}\sum_{t=1}^{T}\hat{u}_{it}\hat{u}_{jt}$,其中 $\hat{u}_{it}$是第 i 个方程的 OLS 残差,$\hat{u}_{jt}$是第 j 个方程的 OLS 残差。VAR 的 BIC 为:

$$\mathrm{BIC}(p)=\ln[\det(\hat{\boldsymbol{\Sigma}}_u)]+k(kp+1)\frac{\ln T}{T} \tag{14.4}$$

其中,$\det(\hat{\boldsymbol{\Sigma}}_u)$是矩阵 $\hat{\boldsymbol{\Sigma}}_u$ 的行列式。计算 AIC 可通过用“2”代替“$\ln T$”项修正公式(14.4)。

在公式(14.4)中,VAR 中 k 个方程的 BIC 表达式扩展了 12.5 节中所给出的单方程表达式。当只有一个方程时,第一项简化为 $\ln[SSR(p)/T]$。公式(14.4)中的第二项是对增加额外回归因子的惩罚;$k(kp+1)$是 VAR 中回归系数的总数(有 k 个方程,每个方程含有一个截距项和 k 个时间序列变量中每个变量的 p 阶滞后)。

VAR 中滞后长度的估计类似于单方程的情形,也使用 BIC:在一组 p 个候选值中,所估计的滞后长度 $\hat{p}$ 是使 BIC(p)最小化的那个 p 值。

将 VAR 用于因果关系分析。到目前为止,我们的讨论都集中在使用 VAR 进行预测上。VAR 模型的另一个应用是分析经济时间序列变量之间的因果关系。事实上,最初正是为了这个目的,VAR 才被经济计量学家和宏观经济学家 Christopher Sims 引入到经济学中。VAR 用于因果关系推断,就是大家所熟知的结构 VAR 建模,之所以称之为“结构”,是因为在这个应用中 VAR 被用来建立基础的经济结构模型。结构 VAR 分析使用本节中所介绍的在预测方面的技术和一些其他的工具。不过,将 VAR 用于预测和结构建模之间最大的概念上的差异是,结构建模需要非常具体的假设,哪些是外生的,哪些不是外生的,这些假设是从经济理论和制度知识中推导出来的。对结构 VAR 的讨论最好是在联立方程系统的估计意义下进行,不过这超出了本书的范围。对于将 VAR 模型用于预测和政策分析方面的介绍,请见 Stock 和 Waston(2001)。关于结构 VAR 建模的额外的数学细节,请见 Hamilton(1994)或 Waston(1994)。

14.1.2 通货膨胀率和失业率的 VAR 模型

作为一个例子,考虑通货膨胀率 Inf_t 和失业率 $Unemp_t$ 的两变量 VAR 模型。和第 12 章一样,我们将通货膨胀率看做是含有随机性趋势的,所以通过计算它的一阶差分 ΔInf_t 进行变换是合适的。

① 这一节用到矩阵,对不需要做较多数学运算的读者来说可以跳过本节。

个季度进行的未来两期的预测值也是过低的。因此，下个季度石油价格的意外上涨意味着上个季度和这个季度的前两期的预测值都太低。由于这个干扰事件，多期回归中的误差项是序列相关的。

如13.4节中所讨论的，如果误差项是序列相关的，那么通常的OLS标准误是不正确的，或更准确地说，它们不是统计推断的可靠基础。因此，在多期回归中必须使用异方差——自相关——一致性(HAC)标准误。因而，本节所报告的多期回归的标准误是纽韦—韦斯特HAC标准误，其中截断参数 m 是根据公式(13.17)来确定的，对这里的数据($T=152$)而言，公式(13.17)得出 $m=4$。随着预测期的变长，误差中序列相关的程度也会增加，一般而言，在前 h 期的回归中，误差项的前 $h-1$ 个自相关系数都是非零的。因此，大的 m 值比公式(13.17)所指示的值更适用于长预测区间的多期回归。

迭代AR预测法：AR(1)　迭代AR预测法使用AR模型将前1期的预测扩展到前两期或更多期的预测。前两期的预测分两步进行计算。第一步，前1期的预测像12.3节中那样来计算。第二步，利用该预期的前1期预测值来计算前两期的预测值。因此，前1期的预测值用作前两期预测的中间步骤。对于更远的期限，这个过程被重复或“迭代”进行。

作为一个例子，考虑 ΔInf_t 的一阶自回归(公式(12.7))，即：

$$\widehat{\Delta Inf_t} = \underset{(0.14)}{0.02} - \underset{(0.11)}{0.21}\Delta Inf_{t-1} \tag{14.9}$$

根据公式(14.9)，使用直到1999:Ⅳ的数据，计算 $\Delta Inf_{2000:\text{II}}$ 的前两期(季)的预测值的第一步，就是根据直到1999:Ⅳ的数据计算 $\Delta Inf_{2000:\text{I}}$ 的前1期(季)的预测值：$\widehat{\Delta Inf}_{2000:\text{I}|1999:\text{IV}} = 0.02-0.21\Delta Inf_{1999:\text{IV}} = 0.02-0.21\times 0.4 = -0.1$。在第二步中，将这个预测值代入公式(14.9)，即 $\widehat{\Delta Inf}_{2000:\text{II}|1999:\text{IV}} = 0.02-0.21\,\widehat{\Delta Inf}_{2000:\text{I}|1999:\text{IV}} = 0.02-0.21\times(-0.1) = 0.0$。因此，根据直到1999年第四季度的信息，这个预测值就是，在2000年第一季度和第二季度之间通货膨胀率将不会发生变化。

迭代AR预测方法：AR(p)　通过将所估计的AR(p)中的 Y_{t-1} 替换为它的前期所做的预测值，可将迭代的AR(1)方法扩展到AR(p)。

例如，考虑12.3节中基于AR(4)模型(公式(12.13))的迭代的前两步通货膨胀预测：

$$\widehat{\Delta Inf_t} = \underset{(0.12)}{0.02} - \underset{(0.10)}{0.21}\Delta Inf_{t-1} - \underset{(0.09)}{0.32}\Delta Inf_{t-2} + \underset{(0.09)}{0.19}\Delta Inf_{t-3} - \underset{(0.10)}{0.04}\Delta Inf_{t-4} \tag{14.10}$$

迭代的前两期预测值，是用预测值 $\widehat{\Delta Inf}_{t|t-1}$ 代替公式(14.10)中的 ΔInf_{t-1} 来计算的。在12.3节中，我们根据直到1999:Ⅳ的数据，使用这个AR(4)模型计算出 $\Delta Inf_{2000:\text{I}}$ 的预测值为 $\widehat{\Delta Inf}_{2000:\text{I}|1999:\text{IV}} = 0.2$。因此，基于AR(4)模型的前两个季度的迭代预测值为 $\widehat{\Delta Inf}_{2000:\text{II}|1999:\text{IV}} = 0.02-0.21\,\widehat{\Delta Inf}_{2000:\text{I}|1999:\text{IV}} - 0.32\Delta Inf_{1999:\text{IV}} + 0.19\Delta Inf_{1999:\text{III}} - 0.04\Delta Inf_{1999:\text{II}} = 0.02-0.21\times 0.2-0.32\times 0.4+0.19\times 0.1-0.04\times 1.1 = -0.2$。根据这个迭代的AR(4)预测值，以直到1999年第四季度的数据为基础，该通货膨胀率在2000年第一季度和第二季度之间将下降0.2个百分点。

多期单变量预测的这两种方法在重要概念14.2中总结。

重要概念 14.2

使用单变量自回归的多期预测

首先,估计多期回归方程:

$$Y_t = \delta_0 + \delta_1 Y_{t-h} + \cdots + \delta_p Y_{t-p-h-1} + u_t \tag{14.11}$$

然后用所估计的系数计算提前 h 期的预测值,进而计算基于 AR(p)的未来 h 期的多期回归预测值(multiperiod regression forecast)。

迭代的 AR 预测(iterated forecast)是分步计算的:先计算往前 1 期的预测值,然后用它计算往前两期的预测值,依此类推。基于 AR(p)模型的往前两期和 3 期的迭代预测是:

$$\hat{Y}_{t|t-2} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t-1|t-2} + \hat{\beta}_2 Y_{t-2} + \hat{\beta}_3 Y_{t-3} + \cdots + \hat{\beta}_p Y_{t-p} \tag{14.12}$$

$$\hat{Y}_{t|t-3} = \hat{\beta}_0 + \hat{\beta}_1 \hat{Y}_{t-1|t-3} + \hat{\beta}_2 Y_{t-2|t-3} + \hat{\beta}_3 Y_{t-3} + \cdots + \hat{\beta}_p Y_{t-p} \tag{14.13}$$

其中,$\hat{\beta}$ 是 AR(p)系数的 OLS 估计值。重复这个过程("迭代"),得到对未来更远期的预测。

14.2.2 多期预测:多变量预测

从单变量模型中得出的多期预测的两种方法同样可用于多变量预测的回归中。

多期回归方法 在一般的多期回归方法中,所有的预测因子都被滞后 h 期,以得到未来 h 期的预测值。

例如,首先通过估计回归方程(14.14)来计算使用 ΔInf_t 和 $Unemp_t$ 中每个四阶滞后的前两个季度的 ΔInf_t 预测值。

$$\begin{aligned}\widehat{\Delta Inf}_{t|t-2} = \underset{(0.40)}{0.27} - \underset{(0.11)}{0.28}\Delta Inf_{t-2} + \underset{(0.10)}{0.15}\Delta Inf_{t-3} - \underset{(0.11)}{0.21}\Delta Inf_{t-4} - \underset{(0.08)}{0.06}\Delta Inf_{t-5} \\ - \underset{(0.46)}{0.21}Unemp_{t-2} + \underset{(0.98)}{0.79}Unemp_{t-3} - \underset{(1.12)}{2.11}Unemp_{t-4} + \underset{(0.56)}{1.49}Unemp_{t-5}\end{aligned} \tag{14.14}$$

将 $\Delta Inf_{1999:Ⅰ}, \cdots, \Delta Inf_{1999:Ⅳ}, Unemp_{1999:Ⅰ}, \cdots, Unemp_{1999:Ⅳ}$ 的值代入公式(14.14),可以计算往前两季的预测值,这得出 $\widehat{\Delta Inf}_{2000:Ⅱ|1999:Ⅳ} = 0.27 - 0.28\Delta Inf_{1999:Ⅳ} + 0.15\Delta Inf_{1999:Ⅲ} - 0.21\Delta Inf_{1999:Ⅱ} - 0.06\Delta Inf_{1999:Ⅰ} - 0.21 Unemp_{1999:Ⅳ} + 0.79 Unemp_{1999:Ⅲ} - 2.11 Unemp_{1999:Ⅱ} + 1.49 Unemp_{1999:Ⅰ} = 0.0$。

通过将公式(14.14)中的所有回归因子滞后 1 个季度,估计那个回归,并计算预测值来计算往前 3 期的 ΔInf_t 预测值,对于未来更远期的预测值也依此类推计算。

迭代 VAR 预测法。迭代的 AR 方法经过一定的修正后就可将其扩展为 VAR 模型,由于 VAR 含有 1 个或多个额外的预测因子,因此,有必要计算所有预测因子的中间预测值。

往前两期的迭代的 VAR 预测(iterated VAR forecast)分两步计算。第一步,和 14.1 节中所讨论的一样,用 VAR 生成该 VAR 中所有变量的往前 1 季的预测值。第二步,用这些预测值代替该 VAR 中的一阶滞后值,也就是说,往前两期的预测值是以往前 1 期的预测值和 VAR 中所设定的额外滞后项为基础的。重复这个过程可得到对未来更远期的迭代 VAR 预测值。

作为一个例子,根据 14.1 节中 ΔInf_t 和 $Unemp_t$ 的 VAR(4)(公式(14.5)和公式(14.6)),以直到 1999:Ⅳ的数据为基础,我们计算 $\Delta Inf_{2000:Ⅱ}$ 的迭代 VAR 预测值。第一步,根据这个 VAR 模型计算往前 1 个季度的预测值 $\widehat{\Delta Inf}_{2000:Ⅱ|1999:Ⅳ}$ 和 $\widehat{Unemp}_{2000:Ⅰ|1999:Ⅳ}$。在 12.3

软件执行起来最方便。

14.3 单整阶数和其他的单位根检验

本节通过强调两个进一步的话题，扩展了12.6节中对随机性趋势的处理方法。首先，一些时间序列的趋势没有被随机游动模型很好地描述，所以我们介绍该模型的推广形式，并讨论了它对这种序列的回归建模的含义。其次，我们继续讨论在时间序列数据中和其他问题中的单位根检验问题，并介绍了另一种单位根检验方法。

14.3.1 其他的趋势模型和单整的阶数

回想一下在12.6节中所介绍的趋势随机游动模型，该模型设定，时期 t 的趋势等于时期 $t-1$ 的趋势加上一个随机误差项。如果 Y_t 服从带漂移项 β_0 的随机游动，那么：

$$Y_t = \beta_0 + Y_{t-1} + u_t \tag{14.18}$$

其中，u_t 是序列无关的。再回忆一下12.6节，如果一个序列含有随机游动趋势，那么它有一个等于1的自回归根。

尽管一个趋势的随机游动模型描述了许多经济时间序列的长期趋势，但有些经济时间序列的趋势比公式(14.18)所隐含的趋势更平滑，即前后期之间变化较小。因此，需要一个不同的模型来描述这种序列的趋势。

平滑趋势的一个模型使得该趋势的一阶差分服从随机游动，即：



$$\Delta Y_t = \beta_0 + \Delta Y_{t-1} + u_t \tag{14.19}$$

其中，u_t 是序列无关的。如果 Y_t 满足公式(14.19)，ΔY_t 服从随机游动，则 $\Delta Y_t - \Delta Y_{t-1}$ 是平稳的。一阶差分的差分 $\Delta Y_t - \Delta Y_{t-1}$ 被称为 Y_t 的二阶差分(second difference)，表示为 $\Delta^2 Y_t = \Delta Y_t - \Delta Y_{t-1}$。在这个术语下，如果 Y_t 满足公式(14.19)，那么它的二阶差分是平稳的。如果一个序列具有公式(14.19)形式的趋势，则该序列的一阶差分有一个等于1的自回归根。

术语"单整阶数"。为了区分这两个趋势模型，引入一些额外的术语是必要的。一个含有随机游动趋势的序列被称为是一阶单整的(integrated of order one)，或 $I(1)$。一个含有公式(14.19)形式的趋势的序列被称为是二阶单整的(integrated of order two)，或 $I(2)$。一个不含随机性趋势的平稳序列被称为是零阶单整的(integrated of order zero)，或 $I(0)$。

在 $I(1)$ 和 $I(2)$ 术语中的单整阶数(order of integration)，是指为了使该序列平稳需要将其差分的次数：如果 Y_t 是 $I(1)$ 的，那么 Y_t 的一阶差分 ΔY_t 是平稳的；如果 Y_t 是 $I(2)$ 的，那么 Y_t 的二阶差分 $\Delta^2 Y_t$ 是平稳的；如果 Y_t 是 $I(0)$，那么 Y_t 是平稳的。

单整阶数在重要概念14.4中总结。

重要概念14.4

单整阶数、差分和平稳性

■如果 Y_t 是一阶单整的，即如果 Y_t 是 $I(1)$ 的，那么 Y_t 有一个单位自回归根，而且它的一阶差分 ΔY_t 是平稳的。

■如果 Y_t 是二阶单整的，即如果 Y_t 是 $I(2)$ 的，那么 ΔY_t 有一个单位自回归根，而且它的二阶差分 $\Delta^2 Y_t$ 是平稳的。

■如果 Y_t 是d阶单整的(intergrated of order d，即 $I(d)$)，那么 Y_t 必须经过 d 次差分来消除它的随机性趋势，也就是说，$\Delta^d Y_t$ 是平稳的。

如何检验一个序列是 $I(2)$ 的还是 $I(1)$ 的？ 如果 Y_t 是 $I(2)$ 的，那么 ΔY_t 就是 $I(1)$ 的，因此 ΔY_t 有一个等于 1 的自回归根。然而，如果 Y_t 是 $I(1)$ 的，那么 ΔY_t 是平稳的。所以，Y_t 是 $I(2)$ 的这一零假设与相应的 Y_t 是 $I(1)$ 的这一备择假设，可以通过检验 ΔY_t 是否含有等于 1 的自回归根来检验。如果 ΔY_t 有单位自回归根的假设被拒绝，那么 Y_t 是 $I(2)$ 的假设被拒绝，支持 Y_t 是 $I(1)$ 的备择假设。

$I(2)$ 和 $I(1)$ 序列的例子：价格水平和通货膨胀率 在第 12 章，我们得出这样的结论，美国通货膨胀率似乎具有随机游动的随机性趋势，也就是说，通货膨胀率是 $I(1)$ 的。如果通货膨胀率是 $I(1)$ 的，那么可以通过一阶差分来消除它的随机性趋势，因此 ΔInf_t 是平稳的。回想一下 12.2 节（公式(12.2)），以年率表示的季度通货膨胀率是价格水平对数的一阶差分乘以 400，即 $Inf_t = 400\Delta p_t$，其中 $p_t = \ln(CPI_t)$。因而，将通货膨胀率看做是 $I(1)$ 的等价于将 Δp_t 看做是 $I(1)$ 的，但是反过来这又等价于将 p_t 看做是 $I(2)$ 的。因此，我们自始至终都是将价格水平的对数看做是 $I(2)$ 的，尽管我们一直没有使用这个术语。

价格水平的对数 p_t 和通货膨胀率绘制在图 14—1 中。价格水平对数（图 14—1(a)）的长期趋势比通货膨胀率（图 14—1(b)）的长期趋势变化得更平滑。价格水平对数的这个平滑的变化趋势是典型的 $I(2)$ 序列。

14.3.2 单位根的 DF-GLS 检验

本节继续讨论 12.6 节中关于单位自回归根检验的问题。我们先介绍单位自回归根的另一种检验方法，即所谓的 DF-GLS 检验。接下来，在一个可选读的数学证明这一节，我们讨论了为什么单位根检验统计量不服从正态分布，即使在大样本条件下。

DF-GLS 检验 ADF 检验是一个被最早提出来的检验单位根零假设的检验，而且是实际中最常用的检验。不过，其他的检验随后被相继提出来，它们中的许多都比 ADF 检验有更高的功效（见重要概念 3.5）。当备择假设为真时，一个具有比 ADF 检验更高功效的检验，更有可能拒绝有单位根这一零假设（序列是平稳的备择假设），因此，一个功效更强的检验有能力更好地将单位 AR 根和数值接近于 1 但小于 1 的根区分开来。

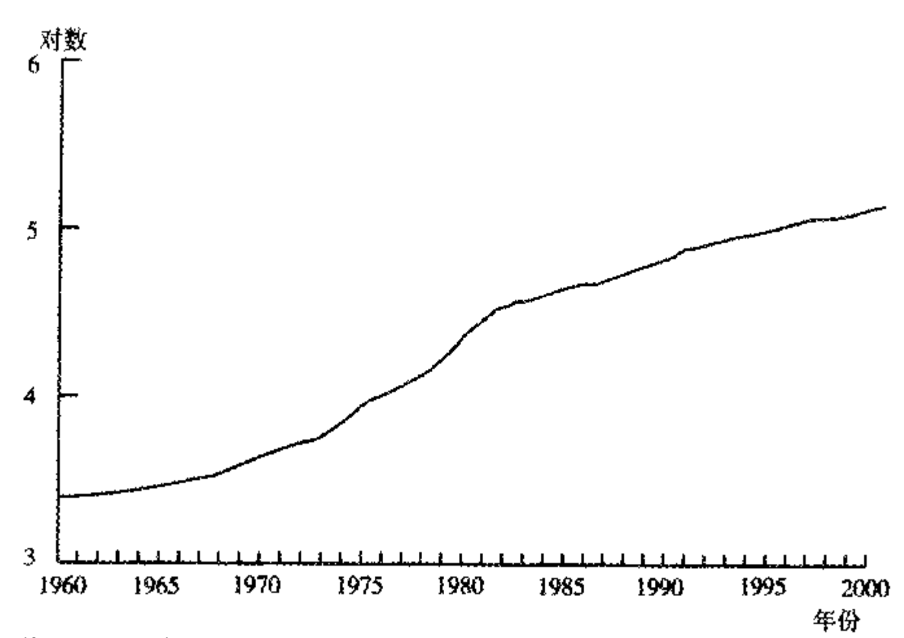
本节讨论由 Elliott，Rothenberg 和 Stock（1996）提出的 DF-GLS 检验（DF-GLS test）。这个检验是针对这样的情况引入的，在零假设下，Y_t 具有随机游动趋势，可能含有漂移项，而在备择假设下，Y_t 围绕一个线性时间趋势是平稳的。

DF-GLS 检验分两步计算。第一步，用广义最小二乘法（GLS，见 13.5 节）估计截距和趋势，通过计算 3 个新的变量 V_t、X_{1t} 和 X_{2t} 来进行 GLS 估计，这里，$V_1 = Y_1, V_t = Y_t - \alpha^* Y_{t-1}, t = 2, \cdots, T, X_{11} = 1, X_{1t} = 1 - \alpha^*, t = 2, \cdots, T, X_{21} = 1, X_{2t} = t - \alpha^*(t-1)$，这里用公式 $\alpha^* = 1 - 13.5/T$ 计算 α^*。然后用 V_t 对 X_{1t} 和 X_{2t} 进行回归，总体回归方程的系数用 OLS 估计，使用 $t = 1, \cdots, T$ 的观测值。

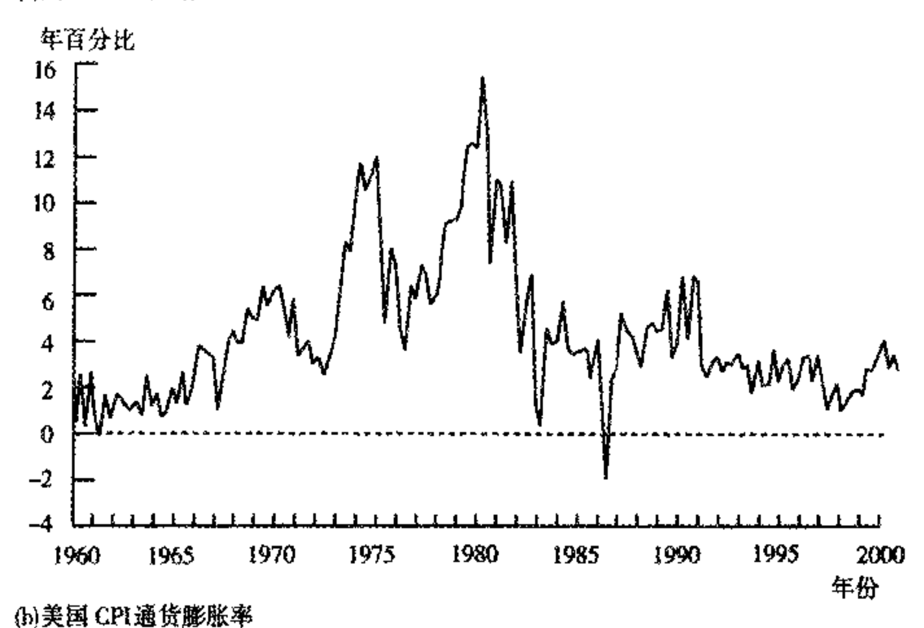
$$V_t = \delta_0 X_{1t} + \delta_1 X_{2t} + e_t \tag{14.20}$$

其中，e_t 是误差项。注意，公式(14.20)的回归中没有截距项，于是，用 OLS 估计量 $\hat{\delta}_0$ 和 $\hat{\delta}_1$ 计算 Y_t 的“去趋势”，即 $Y_t^d = Y_t - (\hat{\delta}_0 + \hat{\delta}_{1t})$。

第二步，用迪基—富勒检验方法来检验 Y_t^d 中的单位自回归根，这里的迪基—富勒回归不包括截距项和时间趋势。也就是说，用 ΔY_t^d 对 Y_{t-1}^d 和 $\Delta Y_{t-1}^d, \cdots, \Delta Y_{t-p}^d$ 进行回归，这里和通常一样，滞后阶数 p 的确定取决于专业知识，或使用 12.5 节中所讨论的基于数据的方法，即 AIC 和 BIC 准则。



(a)美国 CPI 的对数



(b)美国 CPI 通货膨胀率

注：价格对数的趋势（见图 14—1(a)）比通货膨胀率的趋势（见图 14—1(b)）更平滑。

图 14—1 美国价格水平的对数和通货膨胀率（1960—2000）

如果备择假设是"Y_t 是平稳的"，且含有非零的均值，但不含时间趋势，那么前面的步骤就需要修正。具体来讲，用公式 $\alpha^* = 1 - 7/T$ 计算 α^*，省略回归方程（14.20）中的变量 X_{2t}，序列 Y_t^d 被计算为 $Y_t^d = Y_t - \hat{\delta}_0$。

DF-GLS 检验第一步中的 GLS 回归，使得这个检验变得比常规的 ADF 检验更加复杂，但也正是它的这个特点使它改进了区分单位自回归根零假设和 Y_t 是平稳的备择假设的能力。这个改进可能是实质性的。例如，假设 Y_t 实际上是一个平稳的 AR(1)，其自回归系数 $\beta_1 = 0.95$，且有 $T = 200$ 个观测值，计算一个没有时间趋势的单位根检验（即将 t 排除在迪基—富勒回归之外，省略公式（14.20）中的 X_{2t}），那么 ADF 检验在 5% 的显著性水平下正确拒绝零假设的概率约为 31%，而 DF-GLS 检验相应的概率是 75%。

DF-GLS检验的临界值 由于在ADF和DF-GLS检验中确定性项的系数用不同的方法估计,因此这两个检验具有不同的临界值。表14—1给出了DF-GLS检验的临界值。如果DF-GLS检验统计量(第二步回归中Y_{t-1}^d的t统计量)小于该临界值,那么Y_t有单位根的零假设被拒绝。像迪基—富勒检验的临界值一样,适当的临界值取决于使用何种形式的检验,即取决于是否包含时间趋势(X_{2t}是否包含在公式(14.20)中)。

表14—1 DF-GLS检验的临界值

确定性回归因子 (公式(14.20)中的回归因子)	10%	5%	1%
只含截距项(只有X_{1t})	-1.62	-1.95	-2.58
含截距项和时间趋势(X_{1t}和X_{2t})	-2.57	-2.89	-3.48

资料来源:Fuller(1976);Elliott,Rothenberg与Stock(1996,表1)。

在通货膨胀案例中的应用 。当ΔY_t^d的三阶滞后包含在第二阶段的迪基—富勒回归中时,在1962:Ⅰ到1999:Ⅳ期间对CPI通货膨胀率Inf_t所计算的DF-GLS统计量的值是-1.98。这个值刚好小于表14—1中的5%临界值-1.95,所以利用具有3阶滞后的DF-GLS检验导致了在5%的显著水平下拒绝单位根零假设。三阶滞后的选择是基于AIC(在6个滞后中选出最大值)做出的,在此情况下所选择的滞后阶数碰巧与BIC选择的滞后阶数相同。

因为DF-GLS检验有能力更好地区别单位根的零假设和平稳性的备择假设,所以,这个发现的一种解释就是通货膨胀实际上是平稳的,但在12.6节所进行的迪基—富勒检验却没能发现这一点(在5%的水平下)。不过,应该注意的是,DF-GLS检验是否会拒绝零假设,在本例中,对滞后长度的选择是敏感的,因此应该对上述结论保持警惕。如果该检验以四阶滞后为基础,那么它在10%的水平下拒绝了零假设,但在5%的水平下却没有拒绝零假设;如果它以二阶滞后为基础,则它在10%的水平下也没有拒绝零假设。该结论对样本的选择也是敏感的:如果该统计量是在1963:Ⅰ到1999:Ⅳ期间计算的(即正好去掉第一年),则该检验在10%的水平下拒绝零假设,但在5%的水平下不拒绝零假设。因此,总体的描述是相当模棱两可的(因为如在公式(12.34)之后所讨论的,它是以ADF检验为基础的),需要预测者对将通货膨胀建模为$I(1)$的形式还是建模为平稳的形式这个问题做出理性的判断。

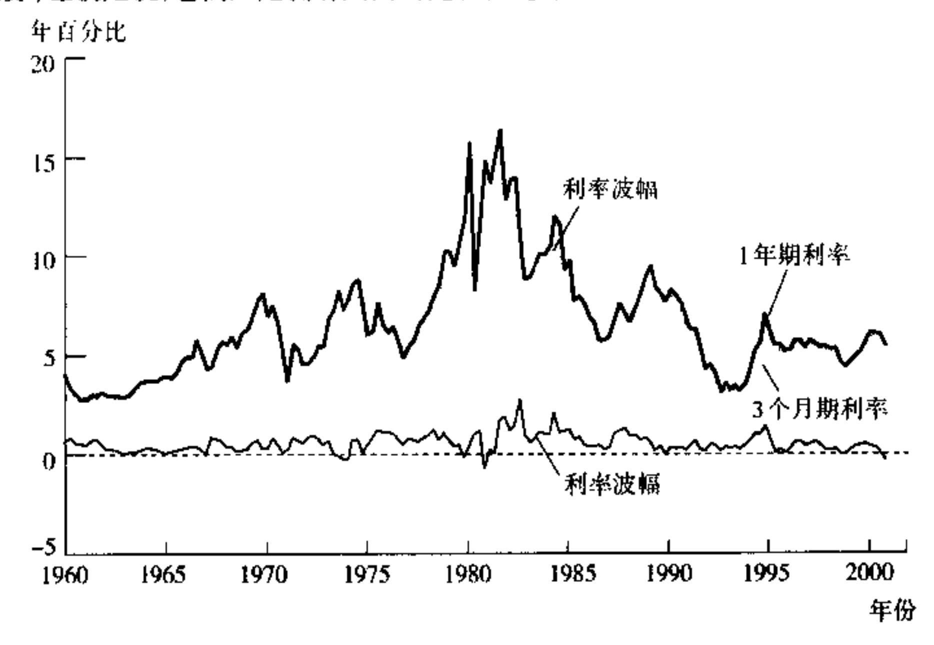
14.3.3 为什么单位根检验具有非正态分布

在12.6节中强调,如果回归因子是非平稳的,那么回归分析强烈依赖的大样本正态分布假设就不适用。在回归包含单位根的零假设下,迪基—富勒回归中的回归因子Y_{t-1}(以及在DF-GLS检验第二步中修正的迪基—富勒回归里的回归因子Y_{t-1}^d)是非平稳的。单位根检验统计量的非正态分布就是这个非平稳性的后果。

为了从数学上更深入地理解这个问题,考虑一个可能是最简单的迪基—富勒回归,在这个回归中,用ΔY_t对单个回归因子Y_{t-1}进行回归,将截距排除在外。用重要概念12.8中的符号,这个回归的OLS估计量为$\hat{\delta}=\sum_{t=1}^{T}Y_{t-1}\Delta Y_t/\sum_{t=1}^{T}Y_{t-1}^2$,于是:

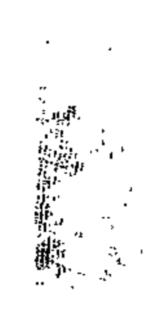
$$T\hat{\delta}=\frac{\frac{1}{T}\sum_{t=1}^{T}Y_{t-1}\Delta Y_t}{\frac{1}{T^2}\sum_{t=1}^{T}Y_{t-1}^2} \tag{14.21}$$

过用一个序列减去另一个序列剔除了每个单独序列中的趋势，所以这两个序列必然具有相同的趋势，也就是说，它们必定拥有共同的随机性趋势。



注：1 年期和 3 个月期利率拥有共同的随机性趋势。这两种利率之间的利差或差分没有表现出趋势。这两种利率看上去是协整的。

图 14—2　1 年期利率、3 个月期利率和利差



两个或两个以上拥有共同随机性趋势的序列被称作是协整的（cointegrated）。协整的正式定义（应归于 Granger，1983）在重要概念 14.5 中给出。在本节中，我们介绍一种检验协整是否存在的方法，讨论与协整变量有关的回归系数的估计，并举例说明协整关系在预测中的应用。讨论一开始主要集中于只有两个变量 X_t 和 Y_t 的情况。

重要概念 14.5

协　整

假设 X_t 和 Y_t 是一阶单整的，如果对于某些系数 θ，$Y_t-\theta X_t$ 是零阶单整的，那么就说 X_t 和 Y_t 是协整的（cointegrated）。系数 θ 被称为协整系数（cointegrating coefficient）。

如果 X_t 和 Y_t 是协整的，那么它们拥有相同的或共同的随机性趋势。计算差值 $Y_t-\theta X_t$ 可以剔除这个共同的随机性趋势。

向量误差修正模型。到目前为止，我们通过计算一阶差分 ΔY_t 已剔除了 $I(1)$ 变量 Y_t 中的随机性趋势。在时间序列回归中，通过使用 ΔY_t 代替 Y_t 就避免了由随机性趋势所引起的问题。不过，如果 X_t 和 Y_t 是协整的，那么剔除趋势的另一种方法是计算差值 $Y_t-\theta X_t$。由于 $Y_t-\theta X_t$ 项是平稳的，因此它也能用在回归分析中。

实际上，如果 X_t 和 Y_t 是协整的，那么可用 VAR 建立 X_t 和 Y_t 的一阶差分模型，并通过将 $Y_t-\theta X_t$ 作为额外的回归因子来扩展回归方程。

$$\Delta Y_t=\beta_{10}+\beta_{11}\Delta Y_{t-1}+\cdots+\beta_{1p}\Delta Y_{t-p}+\gamma_{11}\Delta X_{t-1}+\cdots+\gamma_{1p}\Delta X_{t-p}+\alpha_1(Y_{t-1}-\theta X_{t-1})+u_{1t} \tag{14.24}$$

$$\Delta X_t=\beta_{20}+\beta_{21}\Delta Y_{t-1}+\cdots+\beta_{2p}\Delta Y_{t-p}+\gamma_{21}\Delta X_{t-1}+\cdots+\gamma_{2p}\Delta X_{t-p}+\alpha_2(Y_{t-1}-\theta X_{t-1})+u_{2t} \tag{14.25}$$

$Y_t - \theta X_t$ 项被称为误差修正项(error correction term)。公式(14.24)和公式(14.25)中合并的模型被称为向量误差修正模型(vector error correction model,简写为 VECM)。在 VECM 中,$Y_t - \theta X_t$ 的过去值有助于预测 ΔY_t 和/或 ΔX_t 的未来值。

14.4.2 如何判断两个变量是否是协整的

确定两个变量是否可以合理地建立为协整模型有三种方法:使用专业知识和经济理论;绘出序列图形,并观察它们看上去是否拥有共同的随机性趋势;执行协整统计检验。实践中这三种方法都应该使用。

首先,你必须使用你对这些变量的专业知识来确定协整的存在是否确实是合理的。例如,图 14—2 中的两个利率被所谓的利率期限结构的预期理论联系在一起。根据这个理论,1 年期国债在 1 月 1 日的利率是 90 天短期债券在该年第一季度的利率和在该年第二季度、第三季度及第四季度所发行的未来 90 天短期债券的预期利率的平均值。如果不是这样的,那么投资者就能够预期通过持有 1 年期国债或连续 4 期的 90 天短期债券来赚钱,他们会抬高债券的价格直到预期收益相同。如果这种 90 天利率具有随机游动的随机性趋势,那么这个理论意味着,这个随机性趋势被 1 年期利率所继承,这两种利率之差(即利差)是平稳的。因此,期限结构的预期理论意味着,如果该利率是 $I(1)$ 的,那么它们将是协整的,相应的协整系数 $\theta = 1$(见练习 14.2)。

其次,直观审视序列的图形,有助于确认存在协整的证据。例如,图 14—2 中的两个利率图显示,每个序列看上去都是 $I(1)$ 的,但利差看上去却是 $I(0)$ 的,因此这两个序列看上去是协整的。

再次,迄今为止我们所介绍的单位根检验的方法也可以被扩展到协整检验。这些检验所依据的核心观点是,如果 Y_t 和 X_t 是协整的,协整系数是 θ,那么 $Y_t - \theta X_t$ 是平稳的;否则,$Y_t - \theta X_t$ 是非平稳的(是 $I(1)$ 的)。因此,可以通过检验 $Y_t - \theta X_t$ 有单位根这一零假设来检验 Y_t 和 X_t 不是协整的(即 $Y_t - \theta X_t$ 是 $I(1)$ 的)假设;如果这个假设被拒绝,那么就可以将 Y_t 和 X_t 建模为协整的。这个检验的细节依赖于协整系数 θ 是否是已知的。

当 θ 已知时协整的检验 在一些情况下,专业知识或经济理论可以指明 θ 的值。当 θ 已知时,迪基—富勒和 DF-GLS 单位根检验可被用来检验协整。首先构造序列 $z_t = Y_t - \theta X_t$,然后检验 z_t 有单位自回归根的零假设。

当 θ 未知时协整的检验 如果协整系数 θ 是未知的,那么在检验误差修正模型中的单位根之前必须先估计它。这个初始步骤使得随后的单位根检验有必要使用不同的临界值。

具体来说,第一步,协整系数 θ 用回归的 OLS 估计方法进行估计。

$$Y_t = \alpha + \theta X_t + z_t \tag{14.26}$$

第二步,迪基—富勒 t 检验(含截距项,但无时间趋势)被用于检验这个回归残差 $\hat{z}_t$ 的单位根。这个两步法被称为恩格尔—格兰杰增项的迪基—富勒协整检验或 EG-ADF 检验(test for EG-ADF)(Engle 与 Granger,1987)。

EG-ADF 检验统计量的临界值在表 14—2 中给出。[①] 第一行的临界值只适用于公式(14.26)中有单个回归因子的情形,因此存在两个协整变量(X_t 和 Y_t)。随后各行中的值适用于多个协整变量的情况,它在本节最后讨论。

① 表 14—2 中的临界值取自于 Fuller(1976),Phillips 与 Ouliaris(1990)。根据 Hansen(1992)的建议,表 14—2 中的临界值的选择,可使不论 X_t 和 Y_t 是否含有漂移成分,这些临界值都适用。

表 14—3　　两种利率的单位根和协整检验统计量

序列	ADF 统计量	DF-GLS 统计量
$R90$	-2.96*	-1.88+
$R1yr$	-2.22	-1.37
$R1yr-R90$	-6.31**	-5.59**
$R1yr-1.046R90$	-6.97**	—

注：$R90$ 是以年利率表示的美国 90 天短期债券利率，而 $R1yr$ 是 1 年期美国国债利率。使用 1962：Ⅰ到 1999：Ⅳ期间的季度数据估计回归。在单位根检验统计量回归中，根据 AIC（六阶滞后中最大的一个）来选择滞后阶数。单位根检验统计量在 +10%，*5% 或 **1% 的显著性水平下是显著的。

这个值被强行使用时误差修正项是 $I(0)$ 的（即利差是平稳的），所以原则上不必使用 EG-ADF 检验（θ 是用这个检验估计的）。尽管如此，为了进行说明，我们还是计算了这个检验。EG-ADF 检验的第一步，是用一个变量对另一个变量的 OLS 回归估计 θ，结果是：

$$\widehat{R1yr_t}=0.361+1.046R90_t,\ \bar{R}^2=0.973 \tag{14.28}$$

第二步是计算这个回归残差 $\hat{z}_t$ 的 ADF 统计量。其结果（在表 14—3 的最后一行）小于表 14—2 中给出的 1% 临界值 -3.96，所以，$\hat{z}_t$ 含有单位自回归根的零假设被拒绝。这个统计量也指出了应将这两种利率看做是协整的。注意，公式（14.28）中没有给出标准误，因为如上文所讨论的，协整系数的 OLS 估计量具有一个非正态分布，它的 t 统计量也不是正态分布的，所以给出标准误（HAC 或其他的）会造成误导。



这两种利率的向量误差修正模型。如果 Y_t 和 X_t 是协整的，那么 ΔY_t 和 ΔX_t 的预测，可以通过使用误差修正项的滞后值增加一个 ΔY_t 和 ΔX_t 的 VAR 来改进，即使用公式（14.24）和公式（14.25）中的 VECM 来计算预测值。如果 θ 是已知的，那么 VECM 的未知系数可用 OLS 进行估计，把 $z_{t-1}=Y_{t-1}-\theta X_{t-1}$ 包括进来作为一个增加的回归因子。如果 θ 是未知的，那么 VECM 可将 $\hat{z}_{t-1}$ 作为一个回归因子进行估计，这里的 $\hat{z}_t=Y_t-\hat{\theta}X_t$，其中 $\hat{\theta}$ 是 θ 的一个估计量。

在对这两种利率的应用中，理论建议 $\theta=1$，单位根检验也支持将这两种利率建模为协整系数为 1 的协整关系。因此，我们用理论所建议的值 $\theta=1$ 设定 VECM，即通过将利差的滞后值 $R1yr_{t-1}-R90_{t-1}$ 增加到变量 $\Delta R1yr_t$ 和 $\Delta R90_t$ 的 VAR 模型中来设定 VECM 模型。用一阶差分的二阶滞后来设定，所得的 VECM 为：

$$\begin{aligned}\widehat{\Delta R90_t}=&\underset{(0.17)}{0.14}-\underset{(0.32)}{0.24}\Delta R90_{t-1}-\underset{(0.34)}{0.44}\Delta R90_{t-2}-\underset{(0.39)}{0.01}\Delta R1yr_{t-1}\\&+\underset{(0.27)}{0.15}\Delta R1yr_{t-2}-\underset{(0.27)}{0.18}(R1yr_{t-1}-R90_{t-1})\end{aligned} \tag{14.29}$$

$$\begin{aligned}\widehat{\Delta R1yr_t}=&\underset{(0.16)}{0.36}-\underset{(0.30)}{0.14}\Delta R90_{t-1}-\underset{(0.29)}{0.33}\Delta R90_{t-2}-\underset{(0.35)}{0.11}\Delta R1yr_{t-1}\\&+\underset{(0.25)}{0.10}\Delta R1yr_{t-2}-\underset{(0.24)}{0.52}(R1yr_{t-1}-R90_{t-1})\end{aligned} \tag{14.30}$$

在第一个等式中，没有一个系数在 5% 的水平下是单独显著的，而且滞后的利率一阶差分的系数在 5% 的水平下也不是联合显著的。在第二个等式中，这个滞后的一阶差分的系数不是联合显著的，但滞后利差（误差修正项）系数的估计值为 -0.52，t 统计量为 -2.17，

所以它在5%的水平下在统计上是显著的。尽管利率一阶差分的滞后值对预测未来利率是无用的,但滞后的利差确实有助于预测1年期债券利率的变化。当1年期利率高于90天利率时,据预测,1年期利率将来会下降。

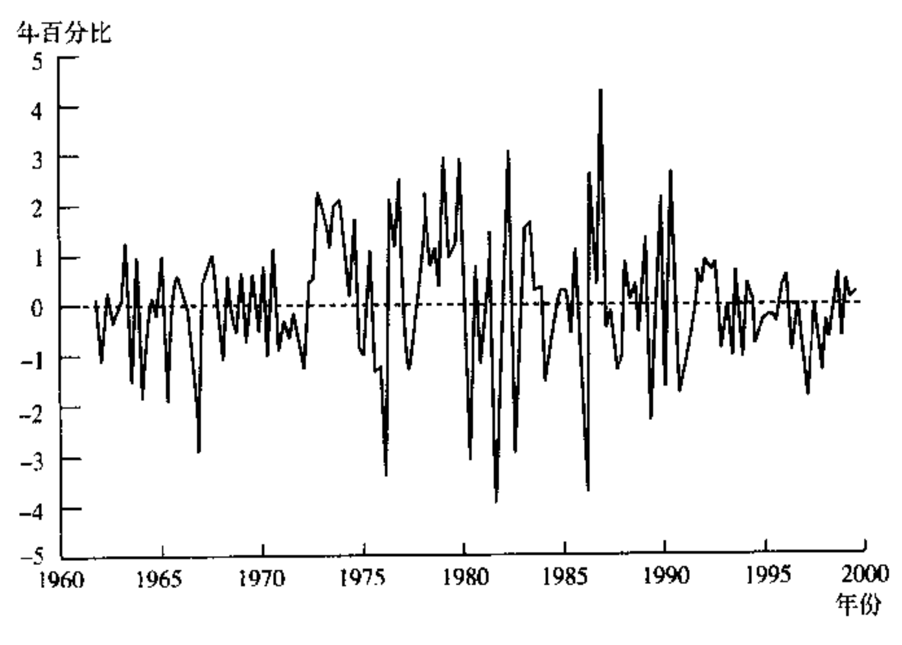
14.5 条件异方差

有时候稳定而有时候不稳定这种现象——即波动以集聚的形式出现——经常出现在许多经济时间序列中。本节提出了量化波动集聚(volatility clustering)的一对模型,也就是众所周知的条件异方差(conditional heteroskedasticity)模型。

14.5.1 波动集聚(volatility clustering)

12.7节有一个奇怪的实证结果:用四阶滞后的菲利浦斯曲线所生成的从1996年到1999年通货膨胀的伪样本外预测的均方根预测误差为0.75个百分点,而生成这些预测值的OLS回归的标准误为1.47。也就是说,样本外误差只是样本内误差的一半!面对这样乐观结果的预测者,当然他或她的客户会感到满意。不过,预测这件事是否也是有时容易有时难呢?20世纪90年代后期正是那些容易的时期之一吗?

根据对绘制在图14—3中的四阶滞后的菲利浦斯曲线残差图的审视,可以得出一些结论:这些残差表现出波动集聚现象。在20世纪70年代后期和80年代早期,绝对预测误差常常超过2个百分点。可是,在20世纪60年代和90年代,绝对预测误差典型地小于1个百分点。



注:菲利浦斯曲线的残差表现出波动集聚现象。在20世纪60年代和90年代波动性相对较低,而在20世纪70年代和80年代波动性相对较高。

图14—3 公式(14.5)中菲利浦斯曲线的残差



波动集聚现象在许多金融时间序列中是司空见惯的。12.2节中所讨论的一个例子显示在图12—2(d)中,这是一个从1990年到1998年NYSE股票综合指数的1 771个日收益图。平均来看,1991年和1998年的绝对日百分比变化比1994年和1995年大。在任何给定

的年份里，一些月份总比另一些月份具有更大的波动性。像菲利浦斯曲线的残差一样，这些价格的百分比变化具有扩展的高波动性时期和扩展的相对稳定时期。

波动集聚可被看做是误差项的方差随时间变化存在集聚现象：如果回归误差在一个时期里具有较小的方差，那么在下一个时期里它的方差也倾向于是较小的。换句话说，波动集聚意味着误差表现出随时间变化的异方差。

14.5.2 自回归条件异方差

波动集聚的两个模型是自回归条件异方差（ARCH）及其扩展形式，即广义 ARCH（GARCH）模型。

ARCH 模型。考虑 ADL(1,1)回归：

$$Y_t = \beta_0 + \beta_1 Y_{t-1} + \gamma_1 X_{t-1} + u_t \tag{14.31}$$

在 ARCH 模型（由经济计量学家 Robert Engle（Engle，1982）提出）中，将误差项 u_t 建立为服从均值为 0、方差为 σ_t^2 的正态分布模型，这里 σ_t^2 取决于 u_t 过去值的平方。具体来说，表示为 ARCH(p)的 p 阶 ARCH 模型是：

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \alpha_2 u_{t-2}^2 + \cdots + \alpha_p u_{t-p}^2 \tag{14.32}$$

其中，$\alpha_0, \alpha_1, \cdots, \alpha_p$ 是未知系数。如果这些系数都是正的，如果近期的平方误差很大，那么 ARCH 模型预测的当前的平方误差在幅度上也将会很大，也即它的方差 σ_t^2 很大。

尽管这里所描述的是公式(14.31)中的 ADL(1,1)模型，但 ARCH 模型可被应用于任何误差项具有条件零均值的时间序列回归模型的误差方差分析，包括更高阶的 ADL 模型、自回归和含有多个预测因子的时间序列回归。

GARCH 模型。由经济计量学家 Timothy Bollerslev（1986）提出的广义 ARCH（GARCH）模型扩展了 ARCH 模型，使 σ_t^2 既依赖于它自身的滞后值，又依赖于平方误差的滞后值。GARCH(p,q)模型为：

$$\sigma_t^2 = \alpha_0 + \alpha_1 u_{t-1}^2 + \cdots + \alpha_p u_{t-p}^2 + \phi_1 \sigma_{t-1}^2 + \cdots + \phi_q \sigma_{t-q}^2 \tag{14.33}$$

其中，$\alpha_0, \alpha_1, \cdots, \alpha_p, \phi_1, \cdots, \phi_q$ 是未知系数。

ARCH 模型类似于一个分布滞后模型，而 GARCH 模型类似于一个 ADL 模型。如在附录 13.2 中所讨论的，ADL 模型（在恰当的时候）能够提供一个比分布滞后模型更简化的动态乘数模型。同样，通过加入 σ_t^2 的滞后项，GARCH 模型能够用比 ARCH 模型少的参数捕捉到方差的缓慢变化。

ARCH 模型和 GARCH 模型一个重要的应用，是测度和预测金融资产收益率随时间变化的波动性，尤其是在很高抽样频率下所观测到的资产，如图 12—2(d)中的日股票收益率。在这类应用中，经常将收益率本身建模为不可预测的，所以公式(14.31)中的回归只包括截距项。

估计和推断。ARCH 模型和 GARCH 模型用极大似然法（见附录 9.2）估计。ARCH 和 GARCH 系数的估计量在大样本条件下服从正态分布，因此，大样本条件下的 t 统计量服从标准正态分布，系数的置信区间可以构造为它的最大似然估计 ±1.96 倍标准误。

14.5.3 应用于通货膨胀预测

公式(14.5)中用 OLS 估计的四阶滞后的菲利浦斯曲线对相同时期的误差项用 GARCH(1,1)模型重新估计，得到：

$$\widehat{\Delta Inf_t} = \underset{(0.33)}{1.29} - \underset{(0.10)}{0.41}\Delta Inf_{t-1} - \underset{(0.09)}{0.31}\Delta Inf_{t-2} + \underset{(0.11)}{0.02}\Delta Inf_{t-3} - \underset{(0.09)}{0.03}\Delta Inf_{t-4}$$

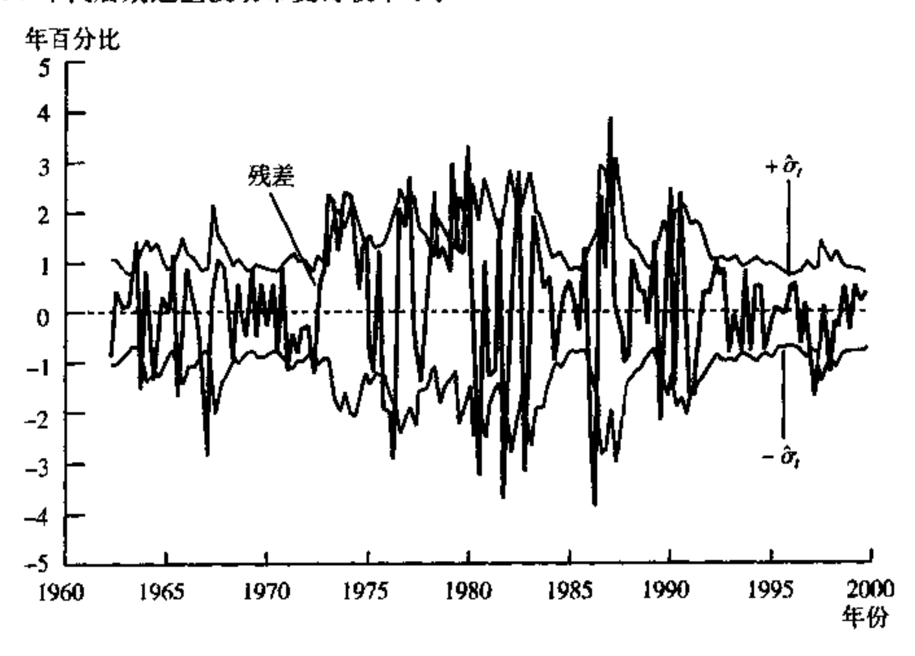
$$- \underset{(0.34)}{2.50}Unemp_{t-1} + \underset{(0.71)}{2.76}Unemp_{t-2} + \underset{(0.81)}{0.15}Unemp_{t-3} - \underset{(0.40)}{0.64}Unemp_{t-4} \quad (14.34)$$

$$\hat{\sigma}_t^2 = \underset{(0.14)}{0.26} + \underset{(0.20)}{0.47}u_{t-1}^2 + \underset{(0.17)}{0.45}\sigma_{t-1}^2 \quad (14.35)$$

GARCH 模型中的两个系数(u_{t-1}^2和 σ_{t-1}^2的系数)在 5% 的显著性水平下在统计上都是单独显著的,而且两个系数都为 0 的联合假设在 5% 的显著性水平下也被拒绝。因此,我们能够在菲利浦斯曲线的误差是条件异方差的备择假设下拒绝它们是同方差的零假设。

用 OLS 估计的 ADL 系数(公式(14.5))和用极大似然估计的带有 GARCH 模型的系数(公式(14.34))稍微不同。如果公式(14.35)中的两个 GARCH 系数确实为 0,那么这两组估计值将会是相同的。不过,这些系数都不是零,因为极大似然估计同时估计了公式(14.34)和公式(14.35)中的系数,所以这两组所估计的 ADL 系数是不同的。

预测方差 $\hat{\sigma}_t^2$ 能够用公式(14.35)中的系数和公式(14.34)中的残差来计算。这些残差,连同基于 GARCH(1,1)模型加减一个预测标准差(即 $\pm\hat{\sigma}_t$)的波动带一起被绘制在图 14—4 中。这些波动带量化了菲利浦斯曲线随时间变化的波动范围。在 20 世纪 80 年代的早期,这些条件标准差的波动带是较宽的,表明了那时在菲利浦斯曲线回归误差中存在相当大的波动性,进而作为所得到的通货膨胀预测有相当大的不确定性。但在 20 世纪 60 年代后期和 90 年代后期这些波动带变得较窄了。



注:用公式(14.35)所计算的 $\pm\hat{\sigma}_t$ 的 GARCH(1,1)波动带在条件方差很小时是窄的,而在条件方差很大时是宽的。$\hat{\sigma}_t$ 在样本的开始期和结尾期较小,此时的预测区间也较窄。

图 14—4 公式(14.34)中菲利浦斯曲线的残差和 GARCH(1,1)波动带



由于掌握了这些条件标准差波动带,我们现在可以回到本节开头的问题上来了:就预测通货膨胀而言,20 世纪 90 年代是个异常的稳定时期吗?所估计的条件方差表明它确实是这样的。例如,在 1993:Ⅳ预测的标准差为 $\hat{\sigma}_{1993:IV} = 0.97$,大大小于公式(14.5)中回归的 OLS 标准

误1.47。实际的伪样本外RMSFE为0.75，仍小于GARCH估计值0.97，但不是小得很多。

14.6 结论

本书的这部分涵盖了时间序列回归中一些最常用的工具和概念。分析经济时间序列的许多其他方面的工具也已为具体的应用设计出来。如果你对了解更多的经济预测感兴趣，请见Enders(1995)和Diebold(2000)的导论性的教科书。对时间序列数据更高级、更现代和更全面的经济计量学处理，请见Hamilton(1994)。

总结

1. 向量自回归将k个时间序列变量的向量建模为每个时间序列依赖于它自身的滞后值和其他$k-1$个序列的滞后值。由VAR模型所生成的每个时间序列的预测都是相互一致的，因为它们是以同样的信息为基础的。

2. 往前两期或更多期的预测，可通过迭代向前一步模型(AR或VAR)或估计一个往前多期的回归来计算。

3. 两个拥有共同随机性趋势的序列是协整的，也就是说，如果Y_t和X_t是$I(1)$，但$Y_t-\theta X_t$是$I(0)$，那么Y_t和X_t是协整的。如果Y_t和X_t是协整的，那么误差修正项$Y_t-\theta X_t$会有助于预测ΔY_t和/或ΔX_t。一个向量误差修正模型就是多包含一个滞后误差修正项的ΔY_t和ΔX_t的VAR模型。

4. 波动集聚——当序列的方差在一些时期高而在另一些时期低时——在经济时间序列中是常见的，尤其是在金融时间序列中。

5. 波动集聚的ARCH模型将回归误差的条件方差表示为近期回归误差平方的函数。GARCH模型将ARCH模型扩展，还包含了滞后的条件方差项。所估计的ARCH模型和GARCH模型生成的预测宽度区间，依赖于最近期间回归残差的波动性。



重要术语

向量自回归(VAR)　多期回归预测值　迭代AR预测　迭代VAR预测　二阶差分　$I(0)$，$I(1)$和$I(2)$　单整阶数　d阶单整($I(d)$)　DF-GLS检验　共同趋势　误差修正项　向量误差修正模型　协整　协整系数　EG-ADF检验　DOLS估计量　波动集聚　条件异方差　ARCH　GARCH

复习概念

14.1　一位宏观经济学家想要对下列宏观经济变量进行预测：GDP、消费、投资、政府采购、出口、进口、短期利率、长期利率和价格通货膨胀率。他有每一个变量从1970年到2001年的季度时间序列数据。他应该估计这些变量的VAR并且将其用于预测吗？为什么？你能建议另一种方法吗？

14.2　假设Y_t服从$\beta_0=0$且$\beta_1=0.7$的平稳AR(1)模型。如果$Y_t=5$，那么Y_{t+2}的预测值(即$Y_{t+2|t}$)是多少？对于$h=30$，$Y_{t+h|t}$是多少？这个$h=30$的预测值对你来说合理吗？

14.3 持久性消费收入理论的一种形式，暗含着实际 GDP(Y)的对数和实际消费(C)的对数是协整的，且协整系数为 1。请解释你如何通过绘制数据图和使用统计检验来研究这个问题的含义。

14.4 考虑 ARCH 模型 $\sigma_t^2 = 1.0 + 0.8u_{t-1}^2$。解释为什么这会导致波动集聚。(提示：当 u_{t-1}^2 异常大时会发生什么?)

14.5 单位根的 DF-GLS 检验比迪基—富勒检验具有更高的功效。为什么要使用一个功效更高的检验?

练习

14.1 假设 Y_t 服从平稳的 AR(1)模型，$Y_t = \beta_0 + \beta_1 Y_{t-1} + u_t$。

*a. 证明：往前 h 期的 Y_t 的预测由 $Y_{t+h|t} = \mu_Y + \beta_1^h (Y_t - \mu_Y)$ 给出，其中，$\mu_Y = \beta_0/(1-\beta_1)$。

b. 假设用 $X_t = \sum_{i=0}^{\infty} \delta^i Y_{t+i|t}$ 将 X_t 与 Y_t 联系在一起，其中 $\delta < 1$。证明：$X_t = \frac{\mu_Y}{1-\delta} + \frac{Y_t - \mu_Y}{1-\beta_1\delta}$。

14.2 利率期限结构的预期理论的一种形式认为，长期利率等于未来短期利率期望值的平均值加上一个 $I(0)$ 的期限溢价。具体来说，设 Rk_t 表示 k 期利率，$R1_t$ 表示 1 期利率，e_t 表示一个 $I(0)$ 的期限溢价，那么 $Rk_t = \frac{1}{k}\sum_{i=1}^{k} R1_{t+i|t} + e_t$，这里 $R1_{t+i|t}$ 是在 t 期所做的 $t+i$ 期 $R1$ 值的预测。假设 $R1_t$ 服从随机游动，即 $R1_t = R1_{t-1} + u_t$。

a. 证明：$Rk_t = R1_t + e_t$。

b. 证明：Rk_t 和 $R1_t$ 是协整的。协整系数是多少?

c. 现假设 $\Delta R1_t = 0.5\Delta R1_{t-1} + u_t$，(b)中的答案会如何变化?

d. 现假设 $R1_t = 0.5R1_{t-1} + u_t$，(b)中的答案会如何变化?

14.3 假设 u_t 服从 ARCH 过程，$\sigma_t^2 = 1.0 + 0.5u_{t-1}^2$。

*a. 设 $E(u_t^2) = \mathrm{var}(u_t)$ 为 u_t 的无条件方差。证明：$\mathrm{var}(u_t) = 2$。

b. 假设以 u_t 的滞后值为条件的 u_t 的分布是 $N(0, \sigma_t^2)$。如果 $u_{t-1} = 0.2$，那么 $\Pr(-3 \le u_t \le 3)$ 是多少? 如果 $u_{t-1} = 2.0$，那么 $\Pr(-3 \le u_t \le 3)$ 又是多少?

14.4 假设 Y_t 满足 AR(p)模型，$Y_t = \beta_0 + \beta_1 Y_{t-1} + \cdots + \beta_p Y_{t-p} + u_t$，其中，$E(u_t | Y_{t-1}, Y_{t-2}, \cdots) = 0$。令 $Y_{t+h|t} = E(Y_{t+h} | Y_t, Y_{t-1}, \cdots)$。证明：对于 $h > p$，$Y_{t+h|t} = \beta_0 + \beta_1 Y_{t-1+h|t} + \cdots + \beta_p Y_{t-p+h|t}$。

14.5 证明公式(14.22)。(提示：用 $\sum_{t=1}^{T} Y_t^2 = \sum_{t=1}^{T} (Y_{t-1} + \Delta Y_t)^2$ 证明 $\sum_{t=1}^{T} Y_t^2 - \sum_{t=1}^{T} Y_{t-1}^2 = 2\sum_{t=1}^{T} Y_{t-1}\Delta Y_t + \sum_{t=1}^{T} \Delta Y_t^2$，并解出 $\sum_{t=1}^{T} Y_{t-1}\Delta Y_t$)

附录 第 14 章中所使用的美国金融数据

像美国联邦储备银行所报告的那样，3 个月期美国债券利率和 1 年期美国国债利率是它们日利率的月度平均值，并转换成年度基准。本章所使用的季度数据是该季度最后一个月的月度平均利率。

第 5 部分

回归分析的经济计量理论

第 15 章

一元线性回归理论

应用经济计量学家学习一些经济计量学理论是很有必要的。通过学习经济计量学理论,可以把统计软件这个"黑盒子"变为一个方便灵活的工具箱,在面对某一项具体的任务时,你可以从中随意选择你所要用的工具。掌握经济计量学理论,可以使你理解这些工具工作的内在道理,以及为了使每个工具发挥出应有的作用,需要满足哪些条件。也许最重要的是,理解了经济计量学理论,可以使你知道在一项具体的应用中,什么时候一项工具的效果是不好的,什么时候你应该寻求不同的经济计量学方法。

本章介绍一元线性回归的经济计量学理论。我们需要处理两个问题。首先,OLS 估计量和 t 统计量的抽样分布的特征是什么?尤其是在什么情况下,第 4 章讲解的统计推断(假设检验和置信区间)是可靠的?其次,在什么情况下,在理论上需要使用 OLS?也就是说,在什么情况下它的抽样分布具有较小的方差?



本书第 1 至第 4 部分描述的经济计量学方法普遍依赖于渐近分布理论,即当样本容量很大时,估计量和检验统计量的抽样分布的理论。全书使用的渐近逼近结论的突出优点是,它们在相当普遍的意义下都是成立的,也就是说,它们在没有限定误差项服从特定分布或没有要求它们是同方差的情况下成立。但是,如果误差项确实具有这样特殊的特征——尤其是,如果误差项是同方差的,甚至更特殊地,误差项是同方差的且服从正态分布的——那么 OLS 估计量将会具有许多我们所希望的理论性质。尽管这些同方差或正态分布的过强的假设在实际应用中可能是不现实的,但是,它们具有理论意义,因为它们允许我们进一步探索 OLS 估计量的效果,更深入地理解 OLS 回归。

本章首先在 15.1 节概述了第 4 章中已给出的一元回归模型,并列出了本章要用到的所有扩展的最小二乘假设。前三个扩展的最小二乘假设是重要概念 4.3 中的假设,它们都是渐近分布理论所需要的。因此,这三个假设在 15.2 节和 15.3 节被用到,而且这里给出了第 4 章中所用到的有关 OLS 估计量和 t 统计量的渐近结论的数学解释。



一般地说,OLS 估计量和 t 统计量的精确分布或有限样本分布是非常复杂的。然而,在一个特殊情况下,这个精确分布是相对简单的,几乎是渐近分布的紧密的映象。这种情况会

重要概念 15.1 中总结了一元回归模型的五个扩展的最小二乘假设。

重要概念 15.1

一元回归的扩展最小二乘假设

一元线性回归模型为：

$$Y_i = \beta_0 + \beta_1 X_i + u_i, i = 1, \cdots, n \tag{15.1}$$

扩展的最小二乘假设是：

1. $E(u_i|X_i) = 0$（条件零均值）；
2. $(X_i, Y_i), i = 1, \cdots, n$，是取自它们联合分布的独立同分布（i. i. d.）抽样；
3. (X_i, u_i) 具有非零的有限的四阶矩；
4. $\mathrm{var}(u_i|X_i) = \sigma_u^2$（同方差性）；
5. 给定 X_i 条件下，u_i 的条件分布是正态分布（正态误差项）。

15.1.2 OLS 估计量

为了易于参考，在这里我们重新写出 β_0 和 β_1 的 OLS 估计量。

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})(Y_i - \overline{Y})}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{15.2}$$

$$\hat{\beta}_0 = \overline{Y} - \hat{\beta}_1 \overline{X} \tag{15.3}$$

公式(15.2)和公式(15.3)的推导见附录4.2。

15.2 渐近分布理论的基本原理

渐近分布理论是样本容量很大时的统计量——估计量、检验统计量，以及置信区间——分布理论。形式上，这个理论涉及刻画一个统计量沿着一个更大的样本序列的抽样分布的行为特征。这个理论是渐近的，因为它刻画了当 $n \to \infty$ 时，在极限状态下统计量的行为特征。

即使样本容量永远不会是无限的，但渐近分布理论在经济计量学和统计学中仍然发挥着重要的作用，原因主要有两个。第一，如果在一项实证应用中所使用的样本观测期数充分大，那么这个渐近极限可以为有限的样本分布提供一个高质量的近似。第二，渐近抽样分布典型地比精确的有限样本分布简单得多，进而在实际中也更易于应用。综合考虑，这两个原因意味着，可靠的且直接的统计推断方法——使用 t 统计量的假设检验和作为 ±1.96 倍标准误的 95% 的置信区间——可以建立在由渐近理论推导出的渐近抽样分布基础之上。

渐近分布理论的两大基石是大数定律和中心极限定理，它们在 2.6 节都进行了介绍。本节首先继续讨论大数定律和中心极限定理，包括对大数定律的证明。然后，我们介绍另外两个分析工具：Slutsky 定理和连续映射定理（Slutsky's theorem and the continuous mapping theorem），它们扩展了大数定律和中心极限定理的用途。例如，我们利用这些工具证明了基于 $\overline{Y}$ 的检验假设 $E(Y) = \mu_0$ 的 t 统计量在零假设下服从标准正态分布。



15.2.1 依概率收敛和大数定律

依概率收敛和大数定律的概念在2.6节中做了介绍。在这里,我们给出了依概率收敛的精确数学定义,接着表述了大数定律的内容,并给予了证明。

一致性和依概率收敛。设 $S_1,S_2,\cdots,S_n,\cdots$ 为随机变量的一个序列。例如,S_n 可能是随机变量 Y 的 n 个观测值的样本均值 $\overline{Y}$。如果对于任意的正常数 δ,当 $n\to\infty$ 时,S_n 在 μ 的 $\pm\delta$ 之内的概率趋向于1,那么称这个随机变量序列 $\{S_n\}$ 依概率收敛(converge in probability)于极限 μ(即 $S_n \xrightarrow{p} \mu$)。也就是说,对于任意的 $\delta>0$,当 $n\to\infty$ 时:

$$S_n \xrightarrow{p} \mu,\text{当且仅当}\ \Pr(|S_n-\mu|\geqslant\delta)\to 0 \tag{15.4}$$

如果 $S_n \xrightarrow{p} \mu$,那么 S_n 被称作是 μ 的一个一致估计量(Consistent estimator)

大数定律。大数定律是指,在 $Y_1,\cdots,Y_n$ 满足一定条件下,样本均值 $\overline{Y}$ 依概率收敛于总体均值。对应于 $Y_1,\cdots,Y_n$ 的不同条件,概率理论家设计出了多种形式的大数定律版本。本书所使用的大数定律版本是,$Y_1,\cdots,Y_n$ 是取自于一个有限方差分布的独立同分布的抽样。这种形式的大数定律(在重要概念2.6中也叙述了)是:

$$\text{如果 } Y_1,\cdots,Y_n \text{ 是独立同分布的},E(Y_i)=\mu_Y,\text{且 } \mathrm{var}(Y_i)<\infty,\text{那么 } \overline{Y}\xrightarrow{p}\mu_Y \tag{15.5}$$

大数定律的思想可以从图2—6中看出来:随着样本容量的增加,$\overline{Y}$ 的抽样分布集中在总体均值 μ_Y 的周围。抽样分布的一个特征是,$\overline{Y}$ 的方差随着样本容量的增加而减少;另一个特征是,随着 n 的增加,$\overline{Y}$ 落在 μ 的 $\pm\delta$ 之外的概率逐渐趋于零。抽样分布的这两个特征实际上是联系在一起的,大数定律的证明就利用了这种联系。

大数定律的证明。$\overline{Y}$ 的方差和 $\overline{Y}$ 落于 μ 的 $\pm\delta$ 之间的概率二者之间的联系,由Chebychev不等式所提供,该不等式的陈述和证明在附录15.2中给出(见公式(15.47))。对于任意正常数 σ,根据 $\overline{Y}$ 写出的Chebychev不等式为:

$$\Pr(|\overline{Y}-\mu_Y|\geqslant\delta)\leqslant \mathrm{var}(\overline{Y})/\delta^2 \tag{15.6}$$

由于 $Y_1,\cdots,Y_n$ 是独立同分布的,且方差为 σ_Y^2,因此,$\mathrm{var}(\overline{Y})=\sigma_Y^2/n$。因而,对于任意 $\delta>0$,$\mathrm{var}(\overline{Y})/\delta^2=\sigma_Y^2/(\delta^2 n)\to 0$。根据表达式(15.6)可得到,对于任意的 $\delta>0$,$\Pr(|\overline{Y}-\mu_Y|\geqslant\delta)\to 0$,这就证明了大数定律。

一些例子。一致性是渐近分布理论中的一个基本概念,因此我们介绍几个总体均值 μ_Y 的一致性和不一致性估计量的例子。设 $Y_i,i=1,\cdots,n$ 是独立同分布的,且具有正的有限方差 σ_Y^2,考虑下面三个 μ_Y 的估计量:(a) $m_a=Y_1$;(b) $m_b=\left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n}a^{i-1}Y_i$,其中 $0<a<1$;(c) $m_c=\overline{Y}+1/n$。这些估计量是一致的吗?

第一个估计量 m_a 就是第一个观测值,因此 $E(m_a)=E(Y_1)=\mu_Y$,从而 m_a 是无偏的。然而,m_a 不是一致的:对于充分小的 δ,$\Pr(|m_a-\mu_Y|\geqslant\delta)=\Pr(|Y_1-\mu_Y|\geqslant\delta)$(因为 $\delta_Y^2>0$)一定是正的,所以,当 $n\to\infty$ 时,$\Pr(|m_a-\mu_Y|\geqslant\delta)$ 并不趋向于零。因此,m_a 不是一致的。这个不一致性并不奇怪,因为 m_a 仅仅使用了一个观测值的信息,随着样本容量的增加,它的分布不可能集中在 μ_Y 的周围。

第二个估计量 m_b 是无偏的,但不是一致的。它是无偏的,是因为:

$$E(m_b)=E\left[\left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n}a^{i-1}Y_i\right]=\left(\frac{1-a^n}{1-a}\right)^{-1}\sum_{i=1}^{n}a^{i-1}\mu_Y=\mu_Y$$



其中，$\sum_{i=1}^{n} a^{i-1} = (1-a^n)\sum_{i=0}^{\infty} a^i = (1-a^n)/(1-a)$。

m_b 的方差是：

$$\mathrm{var}(m_b) = \left(\frac{1-a^n}{1-a}\right)^{-2}\sum_{i=1}^{n} a^{2(i-1)}\sigma_Y^2 = \sigma_Y^2\frac{(1-a^{2n})(1-a)^2}{(1-a^2)(1-a^n)^2} = \sigma_Y^2\frac{(1+a^n)(1-a)}{(1-a^n)(1+a)}$$

当 $n\to\infty$ 时，该方差有极限 $\mathrm{var}(m_b)\to\sigma_Y^2(1-a)/(1+a)$。因此，该估计量的方差不趋向于零，该分布没有集中在 μ_Y 的周围，而且尽管估计量是无偏的，但也不是一致的。我们也许对此感到惊讶，因为所有的观测值都进入了这个估计量，但是大多数观测值得到了非常小的权重（第 i 个观测值的权重与 a^{i-1} 成比例，随着 i 变大，它接近于零），而由于这个原因，如把该估计量看做是一致的，则相当于取消掉了一小部分抽样误差。

第三个估计量 m_c 是有偏的，但却是一致的。它的偏差是 $1/n$：$E(m_c)=E(\overline{Y}+1/n)=\mu_Y+1/n$。但是随着样本容量的增加，偏差趋向于零，且 m_c 是一致的：$\Pr(|m_c-\mu_Y|\geqslant\delta)=\Pr(|\overline{Y}+1/n-\mu_Y|\geqslant\delta)=\Pr[|(\overline{Y}-\mu_Y)+1/n|\geqslant\delta]$。现有 $|(\overline{Y}-\mu_Y)+1/n|\leqslant|\overline{Y}-\mu_Y|+1/n$，所以，如果 $|(\overline{Y}-\mu_Y)+1/n|\geqslant\delta$，必有 $|\overline{Y}-\mu_Y|+1/n\geqslant\delta$，因此，$\Pr[|(\overline{Y}-\mu_Y)+1/n|\geqslant\delta]\leqslant\Pr(|\overline{Y}-\mu_Y|+1/n\geqslant\delta)$，但是 $\Pr[|\overline{Y}-\mu_Y)+1/n|\geqslant\delta]=\Pr(|\overline{Y}-\mu_Y|\geqslant\delta-1/n)\leqslant\sigma_Y^2/[n(\delta-1/n)^2]\to 0$，其中最后一个不等式根据 Chebychev 不等式（见表达式(15.6)，对于 $n>1/\delta$，用 δ 代替 $\delta-1/n$）得出，由此推断 m_c 是一致的。这个例子说明了一个一般性的结论：一个估计量在有限样本下可能是有偏的，但是，如果这个偏差随着样本容量变大而消失，则这个估计量可能仍然是一致的（见练习 15.10）。

15.2.2 中心极限定理与依分布收敛

如果随机变量的一个序列的分布，当 $n\to\infty$ 时，收敛于一个极限，则随机变量的这个序列被称为依分布收敛。中心极限定理是指，在一般条件下，标准化的样本均值依分布收敛于一个正态随机变量。

依分布收敛。设 $F_1,F_2,\cdots,F_n,\cdots$ 是与随机变量的一个序列 $S_1,S_2,\cdots,S_n,\cdots$ 相对应的累积分布函数的序列。例如，S_n 可能是标准化的样本均值 $(\overline{Y}-\mu_Y)/\sigma_{\overline{Y}}$。如果分布函数 $\{F_n\}$ 收敛于 S 的分布 F，那么称该随机变量的序列 S_n 依分布收敛（converge in distribution）于 S（记为 $S_n\xrightarrow{d}S$），也就是说：

$$S_n\xrightarrow{d}S,\text{当且仅当}\lim_{n\to\infty}F_n(t)=F(t) \tag{15.7}$$

这里，该极限在所有的点 t（在该点上极限分布 F 是连续的）都成立，分布 F 被称为 S_n 的渐近分布（asymptotic distribution）。

对比依概率收敛（$\xrightarrow{p}$）和依分布收敛（$\xrightarrow{d}$）这两个概念是很有必要的。如果 $S_n\xrightarrow{p}\mu$，那么随着 n 的增加，S_n 会以很高的概率逼近于 μ。相反，如果 $S_n\xrightarrow{d}S$，那么随着 n 的增加，S_n 的分布将逼近于 S 的分布。

中心极限定理。现在我们用依分布收敛的概念重述中心极限定理。在重要概念 2.7 中，中心极限定理被表述为：如果 $Y_1,\cdots,Y_n$ 是独立同分布的，且 $0<\sigma_Y^2<\infty$，则 $(\overline{Y}-\mu_Y)/\sigma_{\overline{Y}}$ 的渐近分布是 $N(0,1)$，因为 $\sigma_{\overline{Y}}=\sigma_Y/\sqrt{n}$，$(\overline{Y}-\mu_Y)/\sigma_{\overline{Y}}=\sqrt{n}(\overline{Y}-\mu_Y)/\sigma_Y$。因此，中心极限定理可以重述为 $\sqrt{n}(\overline{Y}-\mu_Y)\xrightarrow{d}\sigma_Y Z$，这里 Z 是一个标准正态随机变量。这意味着当 $n\to\infty$ 时，$\sqrt{n}(\overline{Y}-\mu_Y)$ 的分布收敛于 $N(0,\sigma_Y^2)$，习惯上把这个极限简写为：



$$\sqrt{n}(\bar{Y}-\mu_Y) \xrightarrow{d} N(0,\sigma_Y^2) \tag{15.8}$$

也就是说，如果 $Y_1,\cdots,Y_n$ 是独立同分布的，且 $0<\sigma_Y^2<\infty$，那么 $\sqrt{n}(\bar{Y}-\mu_Y)$ 的分布收敛于一个均值为零、方差为 σ_Y^2 的正态分布。

扩展到时间序列数据。2.6 节中所陈述的大数定律和中心极限定理适用于独立同分布的观测值。正如第 12 章中所讨论的，独立同分布假设对时间序列数据是不适合的，从而在将这两个定理应用于时间序列数据之前，还需要对其进行扩展。这些扩展在本质上是技术性的，也就是说结论是相同的——各种形式的大数定律和中心极限定理都适用于时间序列数据——但它们的应用条件是不同的。这在 12.4 节中做了简要的讨论，但是对时间序列变量的渐近分布理论的数学处理超出了本书的范围，有兴趣的读者可以参阅 Hayashi(2000，第 2 章)。

15.2.3 Slutsky 定理和连续映射定理

Slutsky 定理(Slutsky theorem)将依分布收敛和一致性结合起来了。假设 $a_n \xrightarrow{p} a$，其中 a 是常数，且 $S_n \xrightarrow{d} S$，那么：

$$a_n+S_n \xrightarrow{d} a+S,\ a_nS_n \xrightarrow{d} aS,\ 若\ a\neq 0,\ S_n/a_n \xrightarrow{d} S/a \tag{15.9}$$

这三个结论一起被统称为 Slutsky 定理。

连续映射定理(continuous theorem)研究一个随机变量序列 S_n 的一个连续函数 g 的渐近性质。该定理有两部分。第一部分是，如果 S_n 依概率收敛于常数 a，那么 $g(S_n)$ 依概率收敛于 $g(a)$；第二部分是，如果 S_n 依分布收敛于 S，那么 $g(S_n)$ 依分布收敛于 $g(S)$。也就是说，如果 g 是一个连续函数，那么：

$$\begin{aligned}&(\text{i})\ 如果\ S_n \xrightarrow{p} a，那么\ g(S_n) \xrightarrow{p} g(a)\\&(\text{ii})\ 如果\ S_n \xrightarrow{d} S，那么\ g(S_n) \xrightarrow{d} g(S)\end{aligned} \tag{15.10}$$

作为(i)的一个例子，如果 $s_Y^2 \xrightarrow{p} \sigma_Y^2$，那么 $\sqrt{s_Y^2}=s_Y \xrightarrow{p} \sigma_Y$。作为(ii)的一个例子，假设 $S_n \xrightarrow{d} Z$，这里 Z 是一个标准正态随机变量，并设 $g(S_n)=S_n^2$。由于 g 是连续的，因此连续映射定理成立，从而 $g(S_n) \xrightarrow{d} g(Z)$，即 $S_n^2 \xrightarrow{d} Z^2$。换句话说，$S_n^2$ 的分布收敛于一个标准正态随机变量的平方的分布，反过来它又服从 χ_1^2 分布，即 $S_n^2 \xrightarrow{d} \chi_1^2$。

15.2.4 应用于基于样本均值的 t 统计量

现在，我们利用中心极限定理、大数定律和 Slutsky 定理证明：当 $Y_1,\cdots,Y_n$ 是独立同分布的，且 $0<E(Y_i^4)<\infty$ 时，基于 $\bar{Y}$ 的 t 统计量在零假设下服从标准正态分布。

检验基于样本均值 $\bar{Y}$ 的零假设 $E(Y_i)=\mu_0$ 的 t 统计量，已在公式(2.50)中给出。它可写为：



$$t=\frac{\bar{Y}-\mu_0}{s_Y/\sqrt{n}}=\left[\frac{\sqrt{n}(\bar{Y}-\mu_0)}{\sigma_Y}\right]\div\left(\frac{s_Y}{\sigma_Y}\right) \tag{15.11}$$

其中，第二个等式用了一个小的戏法，即通过分子、分母同除以 σ_Y 而得到。

由于 $Y_1,\cdots,Y_n$ 具有二阶矩(这隐含在它们具有四阶矩的假设中，见练习 15.5)，并由于它们是独立同分布的，因此，公式(15.11)中最后一个等式的括号中的第一项服从中心极限

定理:在零假设下,$\sqrt{n}(\overline{Y}-\mu_0)/\sigma_Y \xrightarrow{d} N(0,1)$。又因为 $s_Y^2 \xrightarrow{p} \sigma_Y^2$(在附录 3.3 中已证明),所以 $s_Y^2/\sigma_Y^2 \xrightarrow{p} 1$,公式(15.11)中第二项的比率趋向于 1(见练习 15.4)。因此,公式(15.11)中最后一个等式最后的那个表达式具有与表达式(15.9)中最后一个表达式相同的形式,其中(用公式(15.9)中的记号表示)$S_n = \sqrt{n}(\overline{Y}-\mu_0)/\sigma_Y \xrightarrow{d} N(0,1)$,且 $a_n = s_Y/\sigma_Y \xrightarrow{p} 1$。应用 Slutsky 定理得到 $t \xrightarrow{d} N(0,1)$。

15.3 OLS 估计量和 t 统计量的渐近分布

回想一下第 4 章,在重要概念 4.3 的假设下(重要概念 15.1 中的前三个假设),OLS 估计量 $\hat{\beta}_1$ 是一致的,且 $\sqrt{n}(\hat{\beta}_1-\beta_1)$ 服从渐近正态分布。此外,检验零假设 $\beta_1=\beta_{1,0}$ 的 t 统计量在该零假设下服从渐近标准正态分布。本节归纳了这些结果,并给出了它们的详细证明。

15.3.1 OLS 估计量的一致性和渐近正态性

最初在重要概念 4.4 中陈述的 $\hat{\beta}_1$ 的大样本分布为:

$$\sqrt{n}(\hat{\beta}_1-\beta_1) \xrightarrow{d} N\left(0, \frac{\mathrm{var}(v_i)}{[\mathrm{var}(X_i)]^2}\right) \tag{15.12}$$

其中,$v_i=(X_i-\mu_X)u_i$。附录 4.3 中有该结论的简要证明,但是,省略了一些证明的细节,并且还需要一个近似,那个证明过程没有给出该近似的正式说明。那个证明中所省略的步骤都留作练习 15.3。

表达式(15.12)的一个含义是,$\hat{\beta}_1$ 是一致的估计量(见练习 15.4)。

15.3.2 异方差稳健的标准误的一致性

在前三个最小二乘假设下,$\hat{\beta}_1$ 的异方差稳健的标准误构成了有效统计推断的基础,具体来说:

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} \xrightarrow{p} 1 \tag{15.13}$$

其中,$\sigma^2_{\hat{\beta}_1}=\mathrm{var}(v_i)/\{n[\mathrm{var}(X_i)]^2\}$,而 $\hat{\sigma}^2_{\hat{\beta}_1}$ 是由公式(4.19)所定义的异方差稳健的标准误的平方,即:

$$\hat{\sigma}^2_{\hat{\beta}_1} = \frac{1}{n-2}\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2}{\left[\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\right]^2} \tag{15.14}$$

为了证明表达式(15.13)中的结论,先使用 $\sigma^2_{\hat{\beta}_1}$ 和 $\hat{\sigma}^2_{\hat{\beta}_1}$ 的定义将表达式(15.13)中的比率重写为:

$$\frac{\hat{\sigma}^2_{\hat{\beta}_1}}{\sigma^2_{\hat{\beta}_1}} = \left(\frac{n}{n-2}\right)\left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2}{\mathrm{var}(v_i)}\right] \div \left[\frac{\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2}{\mathrm{var}(X_i)}\right]^2 \tag{15.15}$$

我们需要证明,公式(15.15)右边三个括号中的每一项都依概率收敛于 1。显然,第一项收敛于 1,而根据样本方差的一致性(见附录 3.3),最后一项也依概率收敛于 1。因此,剩

下的工作是只需证明第二项依概率收敛于1，即 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$。

证明 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$ 分两步进行。首先证明 $\frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} \text{var}(v_i)$，然后证明 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$。

现在，假定 X_i 和 u_i 具有八阶矩（即 $E(X_i^8)<\infty$ 和 $E(u_i^8)<\infty$）——一个比第三个最小二乘假设所要求的四阶矩更强的假设。要证明第一步，我们必须证明 $\frac{1}{n}\sum_{i=1}^{n}v_i^2$ 服从表达式(15.5)中的大数定律。要做到这一点，v_i^2 必须是独立同分布的（由第二个最小二乘假设可知），且 $\text{var}(v_i^2)$ 必须是有限的。要证明 $\text{var}(v_i^2)<\infty$，应用 Cauchy-Schwarz 不等式（见附录15.2）：$\text{var}(v_i^2)\leqslant E(v_i^4)=E[(X_i-\mu_X)^4u_i^4]\leqslant\{E[(X_i-\mu_X)^8]E(u_i^8)\}^{1/2}$。因此，如果 X_i 和 u_i 具有八阶矩，那么 v_i^2 具有有限方差，进而满足表达式(15.5)中的大数定律。

第二步是要证明 $\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2 - \frac{1}{n}\sum_{i=1}^{n}v_i^2 \xrightarrow{p} 0$。因为 $v_i=(X_i-\mu_X)u_i$，所以，第二步就是要证明：

$$\frac{1}{n}\sum_{i=1}^{n}[(X_i-\overline{X})^2\hat{u}_i^2-(X_i-\mu_X)^2u_i^2] \xrightarrow{p} 0 \tag{15.16}$$

证明这个结果需要设定 $\hat{u}_i=u_i-(\hat{\beta}_0-\beta_0)-(\hat{\beta}_1-\beta_1)X_i$，将表达式(15.16)括号中的项展开，反复应用 Cauchy-Schwarz 不等式，同时利用 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的一致性。代数细节留作练习15.9。

前面的讨论假定 X_i 和 u_i 具有八阶矩，但这并不是必要的，在较弱的假设——X_i 和 u_i 具有四阶矩（就如在第三个最小二乘假设中所述的）下，就可以证明结论：$\frac{1}{n}\sum_{i=1}^{n}(X_i-\overline{X})^2\hat{u}_i^2 \xrightarrow{p} \text{var}(v_i)$，但是这个证明超出了本书的范围，详见 Hayashi(2000，2.5 节)。

第5部分

15.3.3 异方差稳健的 t 统计量的渐近正态性

我们现在证明，如果最小二乘假设1～3成立，那么在零假设下，检验假设 $\beta_1=\beta_{1,0}$ 的异方差稳健的 t 统计量具有渐近的标准正态分布。

利用异方差稳健的标准误 $SE(\hat{\beta}_1)=\hat{\sigma}_{\hat{\beta}_1}$（定义在公式(15.14)中）所构造的 t 统计量为：

$$t=\frac{\hat{\beta}_1-\beta_{1,0}}{\hat{\sigma}_{\hat{\beta}_1}}=\left[\frac{\sqrt{n}(\hat{\beta}_1-\beta_{1,0})}{\sqrt{n\sigma_{\hat{\beta}_1}^2}}\right]\div\sqrt{\frac{\hat{\sigma}_{\hat{\beta}_1}^2}{\sigma_{\hat{\beta}_1}^2}} \tag{15.17}$$

由表达式(15.12)可知，等式(15.17)的第二个等式之后括号中的项依分布收敛于标准正态随机变量。此外，由于异方差稳健的标准误是一致的（见公式(15.13)），因此 $\sqrt{\hat{\sigma}_{\hat{\beta}_1}^2/\sigma_{\hat{\beta}_1}^2} \xrightarrow{p} 1$（见练习15.4）。利用 Slusky 定理可得 $t \xrightarrow{d} N(0,1)$。



15.4 误差为正态分布时的精确抽样分布

在小样本条件下，OLS 估计量和 t 统计量的分布依赖于回归误差的分布，而且通常是很

15.4.2 同方差惟一的 t 统计量的分布

检验零假设 $\beta_1=\beta_{1,0}$ 的同方差惟一的 t 统计量为：

$$t=\frac{\hat{\beta}_1-\beta_{1,0}}{SE(\hat{\beta}_1)} \tag{15.22}$$

其中，$SE(\hat{\beta}_1)$ 是用 $\hat{\beta}_1$ 的同方差惟一的标准误计算的。将 $SE(\hat{\beta}_1)$ 的公式（见附录4.4中的公式(4.58)）代入公式(15.22)，重新整理得到：

$$t=\frac{\hat{\beta}_1-\beta_{1,0}}{\sqrt{s_u^2/\sum_{i=1}^{n}(X_i-\bar{X})^2}}=\frac{\hat{\beta}_1-\beta_{1,0}}{\sqrt{\sigma_u^2/\sum_{i=1}^{n}(X_i-\bar{X})^2}}\div\sqrt{\frac{s_u^2}{\sigma_u^2}}$$

$$=\frac{(\hat{\beta}_1-\beta_{1,0})/\sigma_{\hat{\beta}_1|X}}{\sqrt{W/(n-2)}} \tag{15.23}$$

其中，$s_u^2=\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2$，$W=\sum_{i=1}^{n}\hat{u}_i^2/\sigma_u^2$。在零假设下，$\hat{\beta}_1$ 具有以 $X_1,\cdots,X_n$ 为条件的 $N(\beta_{1,0},\sigma^2_{\hat{\beta}_1|X})$ 分布，因此，公式(15.23)最后一个表达式中的分子服从标准正态分布 $N(0,1)$。16.4节中证明了 W 服从自由度为 $n-2$ 的卡方分布，而且 W 独立分布于公式(15.23)分子中的标准化的OLS估计量。根据 t 分布的定义（见附录15.1），在五个扩展的最小二乘假设下，同方差惟一的 t 统计量服从自由度为 $n-2$ 的 t 分布。

自由度如何调整才合适？s_u^2 中的自由度调整确保了 s_u^2 是 σ_u^2 的一个无偏估计量，并且当误差服从正态分布时，t 统计量服从学生 t 分布。

由于 $W=\sum_{i=1}^{n}\hat{u}_i^2/\sigma_u^2$ 是一个服从自由度为 $n-2$ 的卡方分布的随机变量，其均值为 $E(W)=n-2$，因此，$E[W/(n-2)]=(n-2)/(n-2)=1$。重新整理 W 定义的公式，我们便有 $E(\frac{1}{n-2}\sum_{i=1}^{n}\hat{u}_i^2)=\sigma_u^2$。因而，自由度修正使 s_u^2 是 σ_u^2 的一个无偏估计量。同样，用 $n-2$ 而不是 n 去除，公式(15.23)最后一个表达式的分母中的项与附录15.1中给出的学生 t 分布随机变量的定义相匹配。也就是说，通过使用自由度调整来计算标准误，当误差服从正态分布时，t 统计量服从学生 t 分布。

15.5 含有同方差误差的OLS估计量的有效性

为什么我们使用OLS估计 β_0 和 β_1 呢？在第4章中，我们说使用OLS估计量的一个原因是，大多数人基本上都是利用OLS进行估计的，这样我们就和其他的实证研究人员"有共同语言"了。尽管如此，这种说法在理论层次上仍是没有说服力的。从理论上讲，选择一种估计量而不是另一种估计量应该以一般性原理或应用某个一致的标准为指导。如3.1节中所讨论的，通常有两个标准可用来选择估计量：一个是估计量应该是无偏的；另一个是它应该具有尽可能小的方差。



本节我们将证明，根据这两个标准，OLS估计量在一定条件下可能是最佳估计量。具体地讲，高斯—马尔可夫定理指出，当回归误差是同方差的时，以 $X_1,\cdots,X_n$ 为条件的OLS估计量在所有的以 $Y_1,\cdots,Y_n$ 为线性条件和所有的条件无偏的（以 $X_1,\cdots,X_n$ 为条件无偏的）估计量中，具有最小的方差。换句话说，OLS估计量是**最佳线性条件无偏估计量**（*best linear*

conditionally unbiased estimator,即 BLUE)。这一结论为使用 OLS 估计量提供了一个重要的理论根据。

我们从陈述高斯—马尔可夫条件开始,这些条件是高斯—马尔可夫定理成立的条件。然后,我们定义线性无偏估计量的类别,证明 OLS 估计量属于此类、而后转到高斯—马尔可夫定理本身上。

15.5.1 高斯—马尔可夫条件

有三个高斯—马尔可夫条件。第一个条件,给定所有回归因子 $X_1,\cdots,X_n$ 的观测值,u_i 具有条件零均值;第二个条件,u_i 是同方差的;第三个条件,以 $X_1,\cdots,X_n$ 为条件的误差项在不同的观测值之间不相关。即三个高斯—马尔可夫条件(Gauss-Markov conditions)是:

$$
\begin{aligned}
&(\mathrm{i})E(u_i|X_1,\cdots,X_n)=0\\
&(\mathrm{ii})\mathrm{var}(u_i|X_1,\cdots,X_n)=\sigma_u^2,0<\sigma_u^2<\infty,i=1,\cdots,n \qquad (15.24)\\
&(\mathrm{iii})E(u_iu_j|X_1,\cdots,X_n)=0,i=1,\cdots,n,j=1,\cdots,n,i\neq j
\end{aligned}
$$

重要概念 15.1 中的前四个最小二乘假设就已隐含了高斯—马尔可夫条件。因为观测值是独立同分布的(假设 2),$E(u_i|X_1,\cdots,X_n)=E(u_i|X_i)$,而且根据假设 1,便有 $E(u_i|X_i)=0$,所以条件(i)成立。同样,根据假设 2,$\mathrm{var}(u_i|X_1,\cdots,X_n)=\mathrm{var}(u_i|X_i)$,又根据假设 4(同方差性),$\mathrm{var}(u_i|X_i)=\sigma_u^2$,它是个常数。假设 3(非零的有限的四阶矩)确保了 $0<\sigma_u^2<\infty$,因此条件(ii)成立。为了证明条件(iii)隐含在前四个最小二乘假设中,要注意 $E(u_iu_j|X_1,\cdots,X_n)=E(u_iu_j|X_i,X_j)$,因为根据假设 2,$(X_i,Y_i)$ 是独立同分布的。假设 2 还隐含着对于 $i\neq j$,$E(u_iu_j|X_i,X_j)=E(u_i|X_i)E(u_j|X_j)$(见练习 15.7),因为对于所有的 i,$E(u_i|X_i)=0$,所以,对于所有的 $i\neq j$,有 $E(u_iu_j|X_1,\cdots,X_n)=0$,从而条件(iii)成立。因此,重要概念 15.1 中的最小二乘假设 1~4 隐含了公式(15.24)中的高斯—马尔可夫条件。

15.5.2 线性条件无偏估计量(linear conditionally unbiased estimators)

线性条件无偏估计量,是由 β_1 的所有以 $X_1,\cdots,X_n$ 为条件、与 $Y_1,\cdots,Y_n$ 为线性函数关系的无偏估计量所组成。OLS 估计量就是一个线性条件无偏估计量。

线性条件无偏估计量的类别。与 $Y_1,\cdots,Y_n$ 为线性关系的估计量是 $Y_1,\cdots,Y_n$ 的加权平均。也就是说,如果 $\tilde{\beta}_1$ 是个线性估计量,那么它可被写为:

$$\tilde{\beta}_1=\sum_{i=1}^{n}a_iY_i \quad (\tilde{\beta}_1\text{ 是线性的}) \qquad (15.25)$$

其中,$a_1,\cdots,a_n$ 是权数,这些权数可能依赖于 $X_1,\cdots,X_n$,也可能依赖于非随机常数,但不会依赖于 $Y_1,\cdots,Y_n$。

如果给定 $X_1,\cdots,X_n$,$\tilde{\beta}_1$ 的条件抽样分布的均值是 β_1,那么估计量 $\tilde{\beta}_1$ 是条件无偏的。也就是说,如果公式(15.26)成立,那么 $\tilde{\beta}_1$ 是条件无偏的。

$$E(\tilde{\beta}_1|X_1,\cdots,X_n)=\beta_1 \quad (\tilde{\beta}_1\text{ 是条件无偏的}) \qquad (15.26)$$

如果估计量 $\tilde{\beta}_1$ 可写成公式(15.25)的形式(它是线性的),而且公式(15.26)成立,那么估计量 $\tilde{\beta}_1$ 是个线性条件无偏估计量。

OLS 估计量 $\hat{\beta}_1$ 是个线性条件无偏估计量。证明 $\hat{\beta}_1$ 是线性的,首先要注意到,因为 $\sum_{i=1}^{n}(X_i-\bar{X})=0$(根据 $\bar{X}$ 的定义),$\sum_{i=1}^{n}(X_i-\bar{X})(Y_i-\bar{Y})=\sum_{i=1}^{n}(X_i-\bar{X})Y_i-\bar{Y}\sum_{i=1}^{n}(X_i-\bar{X})$



$= \sum_{i=1}^{n}(X_i - \overline{X})Y_i$。将这一结果代入公式(15.2)中 $\hat{\beta}_1$ 的公式,得到:

$$\hat{\beta}_1 = \frac{\sum_{i=1}^{n}(X_i - \overline{X})Y_i}{\sum_{i=1}^{n}(X_i - \overline{X})^2} = \sum_{i=1}^{n}\hat{a}_i Y_i,其中,\hat{a}_i = \frac{X_i - \overline{X}}{\sum_{j=1}^{n}(X_j - \overline{X})^2} \tag{15.27}$$

由于公式(15.27)中的权重 $\hat{a}_i, i = 1, \cdots, n$,依赖于 $X_1, \cdots, X_n$,但不依赖于 $Y_1, \cdots, Y_n$,因此,OLS 估计量是个线性估计量。

在高斯—马尔可夫条件下,$\hat{\beta}_1$ 是条件无偏的,且在给定 $X_1, \cdots, X_n$ 条件下,$\hat{\beta}_1$ 的条件分布的方差是:

$$\mathrm{var}(\hat{\beta}_1 \mid X_1 \cdots, X_n) = \sigma^2_{\hat{\beta}_1 | X} = \frac{\sigma^2_u}{\sum_{i=1}^{n}(X_i - \overline{X})^2} \tag{15.28}$$

$\hat{\beta}_1$ 是条件无偏的结论已在公式(15.20)中证明了,公式(15.28)中的方差公式在前面作为公式(15.18)被推导出来。公式(15.18)和公式(15.20)是在重要概念 15.1 中的所有五个最小二乘假设(包括正态分布误差假设)下推导的。但是,从推导过程来看,你可以证明,在等式(15.18)和公式(15.20)中的这些结论,可以在较弱的高斯—马尔可夫条件下成立,尤其是不要求误差项服从正态分布。

15.5.3 高斯—马尔可夫定理

高斯—马尔可夫定理(Gauss-Markov theorem)指出,在公式(15.24)中的高斯—马尔可夫条件下,给定 $X_1, \cdots, X_n$,OLS 估计量 $\hat{\beta}_1$ 在 β_1 的所有线性条件无偏估计量中具有最小的方差,也就是说,OLS 估计量是 BLUE。高斯—马尔可夫定理在重要概念 15.2 中进行了陈述,并在附录 15.3 中进行了证明。

重要概念 15.2

对于 $\hat{\beta}_1$ 的高斯—马尔可夫定理

假设公式(15.24)中的高斯—马尔可夫条件成立,那么 OLS 估计量 $\hat{\beta}_1$ 就是最佳线性无偏估计量(BLUE),也就是说,对所有的线性条件无偏估计量 β_1 面言,$\mathrm{var}(\hat{\beta}_1 | X_1, \cdots, X_n) \leqslant \mathrm{var}(\hat{\beta}_1 | X_1, \cdots, X_n)$。

样本均值是 $E(Y)$ 的线性有效估计量。高斯—马尔可夫定理的一个应用是,当 $Y_1, \cdots, Y_n$ 独立同分布时,样本均值 $\overline{Y}$ 是 $E(Y_i)$ 的最有效的线性估计量。为了弄清楚这一点,考虑一个没有"X"的回归的情况,因此,惟一的回归因子是常数回归因子 $X_{0i} = 1$,那么 OLS 估计量 $\hat{\beta}_0 = \overline{Y}$。由此得出,在高斯—马尔可夫假设下,$\overline{Y}$ 是 BLUE。注意,高斯—马尔可夫的一个要求条件即误差是同方差的,在这里是不重要的,因为这里没有回归因子。因此,如果 $Y_1, \cdots, Y_n$ 是独立同分布的,那么就可以得出 $\overline{Y}$ 是 BLUE。这个结果已在重要概念 3.3 中叙述过了。

当 X 非随机时的高斯—马尔可夫定理。当 X 非随机时,高斯—马尔可夫定理在解释上有一点点变化,即高斯—马尔可夫定理同样适用于非随机回归因子的情形,也即适用于在重复抽样中其值保持不变的那些回归因子。具体来说,如果第二个最小二乘假设被 $X_1, \cdots, X_n$

估计量。在实践中要实际运用 WLS，必须估计函数 h，下面我们就来讨论这个问题。

15.6.2 含有已知函数形式异方差的 WLS

如果异方差具有已知的函数形式，那么异方差函数 h 就可以估计，而且用这个估计函数就可计算 WLS 估计量。

例 1：u 的方差是 X 的二次函数。假设条件方差是已知的，而且是二次函数。

$$\mathrm{var}(u_i|X_i)=\theta_0+\theta_1X_i^2 \tag{15.32}$$

其中，θ_0 和 θ_1 是未知参数，$\theta_0>0$ 且 $\theta_1\geq 0$。

由于 θ_0 和 θ_1 是未知的，因此构造加权变量 $\tilde{Y}_i$，$\tilde{X}_{0i}$ 和 $\tilde{X}_{1i}$ 是不可能的。不过，可以估计 θ_0 和 θ_1，并利用这些估计值计算 $\mathrm{var}(u_i|X_i)$ 的估计值。设 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 为 θ_0 和 θ_1 的估计量，并设 $\widehat{\mathrm{var}}(u_i|X_i)=\hat{\theta}_0+\hat{\theta}_1X_i^2$。定义加权回归因子 $\hat{\tilde{Y}}_i=Y_i/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$，$\hat{\tilde{X}}_{0i}=1/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$，$\hat{\tilde{X}}_{1i}=X_{1i}/\sqrt{\widehat{\mathrm{var}}(u_i|X_i)}$。WLS 估计量就是 $\hat{\tilde{Y}}_i$ 对 $\hat{\tilde{X}}_{0i}$ 和 $\hat{\tilde{X}}_{1i}$ 回归中系数的 OLS 估计量（其中，$\beta_0\hat{\tilde{X}}_{0i}$ 代替了截距 β_0）。

应用这一估计量要求估计条件方差函数，也就是说，估计公式（15.32）中的 θ_0 和 θ_1。一致地估计 θ_0 和 θ_1 的一种方法是用 $\hat{u}_i^2$ 对 X_i^2 进行 OLS 回归，其中，$\hat{u}_i^2$ 是第 i 个 OLS 残差的平方。

假设条件方差具有公式（15.32）的形式，且 $\hat{\theta}_0$ 和 $\hat{\theta}_1$ 是 θ_0 和 θ_1 的一致估计量。在重要概念 15.1 的 1～3 个假设下，再加上由于需要估计 θ_0 和 θ_1 而要求的矩条件，WLS 估计量的渐近分布同 θ_0 和 θ_1 已知时的相同。因此，含有已估计的 θ_0 和 θ_1 的 WLS 估计量与那个不可行的 WLS 具有相同的渐近分布，从这个意义上说，它是渐近的 BLUE。

因为 WLS 的这种方法能够通过估计条件方差函数的未知参数来执行，所以，这种方法有时又被称为**可行的 WLS**（**feasible WLS**）或**估计的 WLS**（**estimated WLS**）。

例 2：方差依赖于第三变量。当条件方差依赖于一个不出现在回归函数中的第三变量 W_i 时，也可以使用 WLS。具体来说，假设搜集了三个变量 $Y_i,X_i,W_i,i=1,\cdots,n$ 的数据；总体回归函数依赖于 X_i 但不依赖于 W_i；条件方差依赖于 W_i，但不依赖于 X_i。也就是说，总体回归函数是 $E(Y_i|X_i,W_i)=\beta_0+\beta_1X_i$，而条件方差是 $\mathrm{var}(u_i|X_i,W_i)=\lambda h(W_i)$，其中 λ 是常数，h 是待估计的函数。

例如，假设一位研究人员对建立某一个州的失业率和该州的经济政策变量（X_i）之间的关系模型感兴趣，但是，所测量的失业率（Y_i）是基于一个抽样调查的失业率，它是真实失业率（Y_i^*）的估计值。因此，用 Y_i 测度 Y_i^* 是有误差的，这里的误差来源于常规的随机抽样误差，从而 $Y_i=Y_i^*+v_i$，其中，v_i 是由调查引起的测量误差。在这个例子中，调查样本容量 W_i 本身并不是真实失业率的一个决定性因素，看上去这似乎是合理的。因此，总体回归函数不依赖于 W_i，即 $E(Y_i^*|X_i,W_i)=\beta_0+\beta_1X_i$。所以，我们有下面两个方程：

$$Y_i^*=\beta_0+\beta_1X_i+u_i^* \tag{15.33}$$

$$Y_i=Y_i^*+v_i \tag{15.34}$$



其中，公式（15.33）建立了州经济政策变量和真实的州失业率之间的关系的模型，而公式（15.34）表示了所测度的失业率 Y_i 与真实失业率 Y_i^* 之间的关系。

在公式（15.33）和公式（15.34）中的模型可以导出一个误差的条件方差依赖于 W_i 但不依赖于 X_i 的总体回归。公式（15.33）中的误差项 u_i^* 代表了被这个回归忽略了的其他因

素，而公式(15.34)中的误差项 v_i 代表了由失业率调查所引起的测量误差。如果 u_i^* 是同方差的，那么 $\text{var}(u_i^* | X_i, W_i) = \sigma_{u^*}^2$ 是常数。可是，调查误差方差反向依赖于调查样本容量 W_i，即 $\text{var}(v_i | X_i, W_i) = a/W_i$，其中，$a$ 是一个常数。由于 v_i 是随机调查误差，我们可以放心地假定它与 u_i^* 无关，因此，$\text{var}(u_i^* + v_i | X_i, W_i) = \sigma_{u^*}^2 + a/W_i$。所以，将公式(15.33)代入公式(15.34)得到一个含有异方差的回归模型：

$$Y_i = \beta_0 + \beta_1 X_i + u_i \tag{15.35}$$

$$\text{var}(u_i | X_i, W_i) = \theta_0 + \theta_1 (1/W_i) \tag{15.36}$$

其中，$u_i = u_i^* + v_i$，$\theta_0 = \sigma_{u^*}^2$，$\theta_1 = a$，$E(u_i | X_i, W_i) = 0$。

如果 θ_0 和 θ_1 是已知的，那么可以使用公式(15.36)中的条件方差函数通过 WLS 估计 β_0 和 β_1。在此例中，θ_0 和 θ_1 是未知的，但是它们可以通过用 OLS 残差的平方（来自公式(15.35)的 OLS 估计）对 $1/W_i$ 进行回归来估计。于是，所估计的条件方差函数能用于构造可行的 WLS 中的权数。

应该强调的是，$E(u_i | X_i, W_i) = 0$ 这个条件是非常关键的；否则，加权误差将含有非零的条件零均值，WLS 也将是不一致的。换句话说，如果 W_i 实际上是 Y_i 的一个决定性因素，那么公式(15.35)就应该是一个同时包含 X_i 和 W_i 的多元回归方程。如果 W_i 和 X_i 是不相关的，W_i 是 Y_i 的一个决定性因素，而且如果 W_i 没有包括在这个回归模型中，那么 OLS 是无偏的，但 WLS 是不一致的。

可行的 WLS 的一般方法。可行的 WLS 一般按照以下四个步骤进行：

1. 用 Y_i 对 X_i 进行 OLS 回归，获得 OLS 残差 $\hat{u}_i$，$i = 1, \cdots, n$。
2. 估计一个条件方差函数 $\text{var}(u_i | X_i)$ 的模型。例如，如果条件方差函数具有方程(15.32)中的形式，就需要用 $\hat{u}_i^2$ 对 X_i^2 回归。一般地说，这一步需要估计一个条件方差 $\text{var}(u_i | X_i)$ 的函数。
3. 利用所估计的函数计算条件方差函数的预测值 $\widehat{\text{var}}(u_i | X_i)$。
4. 利用所估计的条件方差函数平方根的倒数对因变量和回归因子（包括截距）加权。
5. 利用 OLS 估计加权回归的系数，相应所得的估计量就是 WLS 估计量。

回归软件包一般都包括加权最小二乘回归的命令选项，选择该选项则可以自动完成上面的第四步和第五步。



15.6.3 异方差稳健的标准误还是 WLS

处理异方差有两种方法：用 WLS 估计 β_0 和 β_1，或用 OLS 和异方差稳健的标准误估计 β_0 和 β_1。在实践中确定选择何种方法需要权衡每种方法的优缺点。

WLS 的优点是，WLS 比初始回归因子系数的 OLS 估计量更有效，至少是渐近有效的。WLS 的缺点是，它要求知道条件方差函数，并且要估计其参数。如果条件方差函数具有方程(15.32)中的二次函数形式，那么这是很容易做到的。然而在实际中，条件方差函数的函数形式很少是已知的。更严重的是，如果函数形式不正确，那么根据 WLS 回归方法计算的标准误是无效的，因为它们导致不正确的统计推断（检验中有错误的规模）。



使用异方差稳健的标准误的优点是，即使不知道条件方差函数的形式，也可以得到渐近有效的统计推断。另外一个优点是，在现代回归软件包中，都把计算异方差稳健的标准误作为一个可选功能，所以很容易计算，我们无需花费额外的精力考虑异方差的威胁。异方差稳健的标准误的缺点是，OLS 估计量的方差比 WLS 估计量的方差（以真实条件方差函数为基

础)大,至少在渐近意义下如此。

在实际中,$\mathrm{var}(u_i|X_i)$的函数形式几乎是不知道的,这就为在实际应用中使用WLS带来了问题。在一元回归中,这个问题是很难解决的。在多元回归模型中,要确定条件方差的函数形式更加困难。由此,实际使用WLS面临着许多的挑战。相反,在现代统计软件包中,使用异方差稳健的标准误是很简单的,即使在很一般的假设下,所得的统计推断也是可靠的,尤其是,异方差稳健的标准误在不需要设定条件方差的函数形式的情况下就能够使用。综上所述,我们认为,尽管WLS具有理论上的优越性,但是在大多数的实际应用中,异方差稳健的标准误为处理潜在的异方差提供了一种更好的方法。

总结

1. OLS估计量的渐近正态性与异方差稳健的标准误的一致性相结合意味着,如果重要概念15.1中的前三个最小二乘假设成立,那么异方差稳健的t统计量在零假设下服从渐近的标准正态分布。

2. 如果以回归因子为条件的回归误差是独立同分布的且服从正态分布,那么$\hat{\beta}_1$服从以回归因子为条件的精确的正态抽样分布。此外,在零假设下,同方差惟一的t统计量具有精确的学生t_{n-2}抽样分布。

3. 除了重要概念15.1的前三个假设外,如果回归误差是同方差的,那么在β_1所有的线性条件无偏估计量中,OLS估计量$\hat{\beta}_1$是有效的(具有最小方差),也就是说,OLS估计量是最佳线性条件无偏估计量(OLS是BLUE)。

4. 加权最小二乘(WLS)估计量就是应用于加权回归中的OLS,这里,所有的变量都以条件方差$\mathrm{var}(u_i|X_i)$或其估计值的平方根的倒数为权重进行加权。尽管WLS估计量在渐近意义上比OLS估计量更有效,但是,要执行WLS你必须知道条件方差函数的函数形式,而这通常是个苛刻的要求。



重要术语

依概率收敛　一致估计量　依分布收敛　渐近分布　Slutsky定理　连续映射定理　BLUE　高斯—马尔可夫条件　高斯—马尔可夫定理　加权最小二乘法(WLS)　WLS估计量　不可行的WLS　可行的WLS　正态概率密度函数　二元正态概率密度函数

复习概念

15.1　假设重要概念15.1中的假设4成立,但是你在大样本条件下使用异方差稳健的标准误构造了β_1的95%的置信区间。这个置信区间会是渐近有效的吗?也就是说,对于很大的n,它在95%的所有重复抽样中包含β_1的真实值吗?反过来,假设重要概念15.1中的假设4不成立,但是你在大样本条件下用同方差惟一的标准误公式构造了β_1的95%的置信区间。这个置信区间是渐近有效的吗?



15.2　假设A_n是一个依概率收敛于3的随机变量,B_n是一个依分布收敛于标准正态分布的随机变量。A_nB_n的渐近分布是什么?使用这个渐近分布计算$\Pr(A_nB_n)<2$的近

似值。

15.3 假设 Y 和 X 以下列回归的方式联系在一起：$Y=1.0+2.0X+u$。一位研究人员拥有 Y 和 X 的观测值，其中 $0\leqslant X\leqslant 20$。对于 $0\leqslant X\leqslant 10$，条件方差 $\text{var}(u_i|X_i=x)=1$，而对于 $10<X\leqslant 20$，$\text{var}(u_i|X_i=x)=16$。画出观测值 (X_i,Y_i)，$i=1,\cdots,n$ 的假设的散点图。WLS 会给 $X\leqslant 10$ 或给 $X>10$ 赋予较高的权重吗？为什么？

15.4 在上面的问题中，该研究人员决定不用 WLS，而是只用 $X\leqslant 10$ 的观测值计算 OLS 估计量，然后再用 $X>10$ 的观测值计算 OLS 估计量，最后取估计量的两个 OLS 的平均。这种做法比 WLS 好吗？

练习

15.1 考虑不含截距项的回归模型 $Y_i=\beta_1X_i+u_i$（因此，截距 β_0 的真实值为零）。

a. 对这个有约束的回归模型 $Y_i=\beta_1X_i+u_i$，推导 β_1 的最小二乘估计量。这个估计量被称为 β_1 的有约束的最小二乘估计量（$\hat{\beta}_1^{RLS}$），因为它是在一个约束条件下估计的（这里 $\beta_0=0$）。

b. 在重要概念 15.1 的假设 1～3 下，推导 $\hat{\beta}_1^{RLS}$ 的渐近分布。

c. 证明：$\hat{\beta}_1^{RLS}$ 是线性的（见公式(15.25)），并且在重要概念 15.1 的假设 1～2 下，它是条件无偏的（见公式(15.26)）。

d. 在高斯—马尔可夫条件下，推导 $\hat{\beta}_1^{RLS}$ 的条件方差。

e. 在高斯—马尔可夫条件下，比较(d)中 $\hat{\beta}_1^{RLS}$ 的条件方差和 OLS 估计量 $\hat{\beta}_1$（从包含截距的回归中得到的）的条件方差。哪一个估计量是更有效的？使用这里的条件方差的公式来解释。

f. 在重要概念 15.1 的假设 1～5 下，推导 $\hat{\beta}_1^{RLS}$ 的精确抽样分布。

g. 现在考虑估计量 $\tilde{\beta}_1=\sum_{i=1}^{n}Y_i/\sum_{i=1}^{n}X_i$。在高斯—马尔可夫条件下，推导 $\text{var}(\tilde{\beta}_1|X_1,\cdots,X_n)-\text{var}(\hat{\beta}_1^{RLS}|X_1,\cdots,X_n)$ 的表达式，并使用这个表达式证明 $\text{var}(\tilde{\beta}_1|X_1,\cdots,X_n)\geqslant\text{var}(\hat{\beta}_1^{RLS}|X_1,\cdots,X_n)$。

*15.2 假设 (X_i,Y_i) 是独立同分布的，且具有有限四阶矩。证明：样本协方差是总体协方差的一个一致估计量，即 $s_{XY}\xrightarrow{p}\sigma_{XY}$，其中，$s_{XY}$ 按公式(3.22)中的定义。（提示：使用附录 3.3 中的方法和 Cauchy-Schwarz 不等式）



15.3 本题补充了附录 4.3 中给出的 $\hat{\beta}_1$ 渐近分布推导的详细过程。

a. 使用公式(15.19)推导下式。

$$\sqrt{n}(\hat{\beta}_1-\beta_1)=\frac{\sqrt{\frac{1}{n}}\sum_{i=1}^{n}v_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2}-\frac{(\bar{X}-\mu_X)\sqrt{\frac{1}{n}}\sum_{i=1}^{n}u_i}{\frac{1}{n}\sum_{i=1}^{n}(X_i-\bar{X})^2} \tag{15.37}$$

其中，$v_i=(X_i-\mu_X)u_i$。



b. 使用中心极限定理、大数定律和 Slutsky 定理证明，公式(15.37)中最后一项依概率收敛于 0。

c. 使用 Cauchy-Schwarz 不等式和重要概念 15.1 中的第三个最小二乘假设证明，

不过由于 Y 是连续的，它的概率和期望值的数学表达式涉及积分，而不是适用于离散随机变量的求和。

设 f_Y 表示 Y 的概率密度函数。因为概率不能是负数，所以对于所有的 y 有 $f_Y(y) \geqslant 0$。Y 落于 a 和 b（这里 $a < b$）之间的概率为：

$$\Pr(a \leqslant Y \leqslant b) = \int_a^b f_Y(y)\,dy \tag{15.38}$$

因为 Y 必定取实数轴上的某个值，所以 $\Pr(-\infty \leqslant Y \leqslant \infty) = 1$，这意味着 $\int_{-\infty}^{\infty} f_Y(y)\,dy = 1$。

与离散随机变量一样，连续随机变量的期望值和矩是它们值的概率加权平均值，只是求和（例如，公式(2.4)中的求和）被积分所取代。因此，Y 的期望值为：

$$E(Y) = \mu_Y = \int y f_Y(y)\,dy \tag{15.39}$$

其中，积分的范围是使 f_Y 取非零值的 Y 的所有值的集合，方差是 $(Y-\mu_Y)^2$ 的期望值，随机变量的 r 阶矩是 Y^r 的期望值，因此：

$$\mathrm{var}(Y) = E(Y-\mu_Y)^2 = \int (y-\mu_Y)^2 f_Y(y)\,dy \tag{15.40}$$

$$E(Y^r) = \int y^r f_Y(y)\,dy \tag{15.41}$$

正态分布

单变量正态分布。一个正态分布随机变量的概率密度函数（正态 p. d. f.（normal p. d. f.））是：

$$f_Y(y) = \frac{1}{\sigma_Y\sqrt{2\pi}}\exp\left[-\frac{1}{2}\left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right] \tag{15.42}$$

其中，$\exp(x)$ 是 x 的指数函数。公式(15.42)中的因子 $1/(\sigma_Y\sqrt{2\pi})$ 确保了 $\Pr(-\infty \leqslant Y \leqslant \infty) = \int_{-\infty}^{\infty} f_Y(y)\,dy = 1$。

当 $\mu_Y = 0, \sigma_Y^2 = 1$ 时，正态分布被称为标准正态分布。标准正态的 p. d. f. 记为 ϕ，标准正态的 c. d. f. 记为 Φ。因此，标准正态密度为 $\phi(y) = \exp(-y^2/2)/\sqrt{2\pi}$，且 $\Phi(y) = \int_{-\infty}^{y} \phi(s)\,ds$。

双变量正态分布。两个随机变量 X 和 Y 的二元正态 p. d. f.（bivariate normal p. d. f.）为：

$$g_{X,Y}(x,y) = \frac{1}{2\pi\sigma_Y\sigma_Y\sqrt{1-\rho_{XY}^2}} \times \exp\left\{\frac{1}{-2(1-\rho_{XY}^2)}\left[\left(\frac{x-\mu_X}{\sigma_X}\right)^2 - 2\rho_{XY}\left(\frac{x-\mu_X}{\sigma_X}\right)\left(\frac{y-\mu_Y}{\sigma_Y}\right) + \left(\frac{y-\mu_Y}{\sigma_Y}\right)^2\right]\right\} \tag{15.43}$$

其中，ρ_{XY} 是 X 和 Y 之间的相关系数。

当 X 与 Y 不相关（$\rho_{XY} = 0$）时，$g_{X,Y}(x,y) = f_X(x)f_Y(y)$，其中，$f$ 为公式(15.42)中给出的正态密度。这证明了如果 X 和 Y 服从联合正态分布，且是不相关的，那么它们是独立分布的。这是正态分布的特殊性质，其他的分布通常不具有这个性质。

多元正态分布是对双变量正态分布的扩展，以便处理两个以上的随机变量。这种分布利用矩阵表述最方便，附录 16.1 中将给出其表述。



条件正态分布。设 X 和 Y 服从联合正态分布,那么给定 X 下 Y 的条件分布是正态的,均值为 $\mu_{Y|X}=\mu_Y+(\sigma_{XY}/\sigma_X^2)(X-\mu_X)$,方差为 $\sigma_{Y|X}^2=(1-\rho_{XY}^2)\sigma_Y^2$。在条件 $X=x$ 下,这个条件分布的均值是 x 的线性函数,但方差不依赖于 x。

有关的分布

卡方分布。设 $Z_1,Z_2,\cdots,Z_n$ 是 n 个独立同分布的标准正态随机变量,那么随机变量

$$W=\sum_{i=1}^{n}Z_i^2 \tag{15.44}$$

服从自由度为 n 的卡方分布。这个分布记为χ_n^2。因为 $E(Z_i^2)=1$,所以 $E(W)=n$。

学生 t 分布。设 Z 服从标准正态分布,W 服从χ_m^2 分布,且 Z 和 W 是独立分布的,那么随机变量

$$t=\frac{Z}{\sqrt{W/m}} \tag{15.45}$$

服从自由度为 m 的学生 t 分布,记为 t_m。t_∞ 分布就是标准正态分布。

F 分布。设 W_1,W_2 分别是自由度为 n_1,n_2 的卡方分布的独立随机变量,那么随机变量

$$F=\frac{W_1/n_1}{W_2/n_2} \tag{15.46}$$

服从自由度为(n_1,n_2)的 F 分布。这个分布记为 F_{n_1,n_2}。

F 分布依赖于分子自由度 n_1 和分母自由度 n_2。随着分母自由度的变大,F_{n_1,n_2}分布很好地被用 n_1 除的$\chi_{n_1}^2$分布近似。在极限状态下,$F_{n_1,\infty}$分布与用 n_1 除的$\chi_{n_1}^2$分布相同,即它与$\chi_{n_1}^2/n_1$ 分布相同。

附录 15.2 两个不等式

本附录叙述并证明了 Chebychev 不等式和 Cauchy-Schwarz 不等式。

Chebychev 不等式

Chebychev 不等式使用随机变量 V 的方差确定 V 超出其均值 $\pm\delta$ 范围的概率,其中 δ 是个正的常数。

$$\Pr(|V-\mu_V|\geqslant\delta)\leqslant\mathrm{var}(V)/\delta^2 \qquad \text{(Chebychev 不等式)} \tag{15.47}$$

要证明不等式(15.47),设 $W=V-\mu_V$,f 是 W 的 p. d. f. ,δ 为任意正数,现有:



$$\begin{aligned}E(W^2)&=\int_{-\infty}^{\infty}w^2f(w)\,dw\\&=\int_{-\infty}^{-\delta}w^2f(w)\,dw+\int_{-\delta}^{\delta}w^2f(w)\,dw+\int_{\delta}^{\infty}w^2f(w)\,dw\\&\geqslant\int_{-\infty}^{-\delta}w^2f(w)\,dw+\int_{\delta}^{\infty}w^2f(w)\,dw\\&\geqslant\delta^2\left[\int_{-\infty}^{-\delta}f(w)\,dw+\int_{\delta}^{\infty}f(w)\,dw\right]\\&=\delta^2\Pr(|W|\geqslant\delta)\end{aligned} \tag{15.48}$$

其中,第一个等式是 $E(W^2)$的定义;第二个等式成立是因为积分区间分割了整个实数轴;第一个不等式成立是因为被去掉的项是非负的;第二个不等式成立是因为在积分区间上 $w^2\geqslant$

δ^2；而根据 $\Pr(|W|\geq\delta)$ 的定义，最后一个等式成立。将 $W=V-\mu_V$ 代入到最后一个表达式中，注意 $E(W^2)=E[(V-\mu_V)^2]=\text{var}(V)$，并重新整理，便得到不等式(15.47)。如果 V 是离散的，那么用求和代替积分，这个证明仍适用。

Cauchy-Schwarz 不等式

Cauchy-Schwarz 不等式是考虑到了非零均值的相关系数不等式 $|\rho_{XY}|\leq 1$ 的扩展。Cauchy-Schwarz 不等式为：

$$|E(XY)|\leq\sqrt{E(X^2)E(Y^2)}\quad\text{(Cauchy-Schwarz 不等式)}\tag{15.49}$$

不等式(15.49)的证明类似于附录2.1中的相关系数不等式的证明。设 $W=Y+bX$，其中 b 是常数，那么 $E(W^2)=E(Y^2)+2bE(XY)+b^2E(X^2)$。现在设 $b=-E(XY)/E(X^2)$，于是(经简化后)该表达式变为 $E(W^2)=E(Y^2)-[E(XY)]^2/E(X^2)$。由于 $E(W^2)\geq 0$(因为 $W^2\geq 0$)，因此必有 $[E(XY)]^2\leq E(X^2)E(Y^2)$。于是，两边开平方就得到了 Cauchy-Schwarz 不等式。

附录 15.3　高斯—马尔可夫定理的证明

我们先来推导对于所有的线性无偏估计量都成立的一些事实，即对于所有满足公式(15.25)和公式(15.26)的估计量 $\tilde{\beta}_1$ 的一些事实。将 $Y_i=\beta_0+\beta_1X_i+u_i$ 代入 $\tilde{\beta}_1=\sum_{i=1}^{n}a_iY_i$，并合并同类项，得：

$$\tilde{\beta}_1=\beta_0\left(\sum_{i=1}^{n}a_i\right)+\beta_1\left(\sum_{i=1}^{n}a_iX_i\right)+\sum_{i=1}^{n}a_iu_i\tag{15.50}$$

根据第一个高斯—马尔可夫条件，$E\left(\sum_{i=1}^{n}a_iu_i\mid X_1,\cdots,X_n\right)=\sum_{i=1}^{n}a_iE(u_i\mid X_1,\cdots,X_n)=0$，因此，对公式(15.50)的两边同时取条件期望，便得到 $E(\tilde{\beta}_1\mid X_1,\cdots,X_n)=\beta_0\left(\sum_{i=1}^{n}a_i\right)+\beta_1\left(\sum_{i=1}^{n}a_iX_i\right)$。由于根据假设 $\tilde{\beta}_1$ 是条件无偏的，因此，必然有 $\beta_0\left(\sum_{i=1}^{n}a_i\right)+\beta_1\left(\sum_{i=1}^{n}a_iX_i\right)=\beta_1$。但是，要使该等式对所有的 β_0 和 β_1 都成立，且使 $\tilde{\beta}_1$ 是条件无偏的，必须有：



$$\sum_{i=1}^{n}a_i=0,\ \sum_{i=1}^{n}a_iX_i=1\tag{15.51}$$

在高斯—马尔可夫条件下，以 $X_1,\cdots,X_n$ 为条件的 $\tilde{\beta}_1$ 的方差具有简单的形式。将公式(15.51)代入公式(15.50)，得 $\tilde{\beta}_1-\beta_1=\sum_{i=1}^{n}a_iu_i$。因此，$\text{var}(\tilde{\beta}_1\mid X_1,\cdots,X_n)=\text{var}\left(\sum_{i=1}^{n}a_iu_i\mid X_1,\cdots,X_n\right)=\sum_{i=1}^{n}\sum_{j=1}^{n}a_ia_j\text{cov}(u_i,u_j\mid X_1,\cdots,X_n)$，应用第二个和第三个高斯—马尔可夫条件，双重求和中的交叉项消失，从而，条件方差的表达式简化为：

$$\text{var}(\tilde{\beta}_1\mid X_1,\cdots,X_n)=\sigma_u^2\sum_{i=1}^{n}a_i^2\tag{15.52}$$



注意，公式(15.51)和公式(15.52)适用于公式(15.27)中给出的权重 $a_i=\hat{a}_i$ 时的 $\hat{\beta}_1$。

现在我们证明，公式(15.51)中的两个约束条件和公式(15.52)中的条件方差表达式意味着 $\tilde{\beta}_1$ 的条件方差大于 $\hat{\beta}_1$ 的条件方差，除非 $\tilde{\beta}_1=\hat{\beta}_1$。设 $a_i=\hat{a}_i+d_i$，则 $\sum_{i=1}^{n}a_i^2=\sum_{i=1}^{n}(\hat{a}_i^2+$

第 16 章

多元回归理论

本章对多元回归分析理论做了一个基本的介绍。本章有三个目标。第一个目标是给出多元回归模型的矩阵形式表示，它生成了 OLS 估计量和检验统计量的简洁公式。第二个目标是刻画大样本条件下(使用渐近理论)OLS 估计量的抽样分布和刻画小样本条件下(如果误差是同方差的且服从正态分布时)OLS 估计量的抽样分布。第三个目标是研究多元回归模型系数的有效估计理论，并介绍广义最小二乘法(GLS)。GLS 方法是当误差是异方差的和/或误差在不同的观测值之间相关时有效地估计回归系数的一种方法。

本章首先在 16.1 节列出用矩阵形式表示的多元回归模型和 OLS 估计量。这一节还给出了多元回归模型扩展的最小二乘假设。前四个假设与重要概念 15.4 中的最小二乘假设相同，这些假设构成了用来证明第 5 章中所描述的方法是合理的那个渐近分布的基础。其余两个扩展的最小二乘假设更强，允许我们可以更加详细、深入地探讨多元回归模型中 OLS 估计量的理论性质。

接下来的三节考察了 OLS 估计量和检验统计量的抽样分布。16.2 节给出了在重要概念 5.4 的最小二乘假设下 OLS 估计量和 t 统计量的渐近分布。16.3 节统一并归纳了 5.7 节和 5.8 节中所给出的涉及多个回归系数的假设检验，并且给出了相应的 F 统计量的渐近分布。在 16.4 节，我们研究了在误差项是同方差的且服从正态分布的特殊情况下，OLS 估计量和检验统计量的精确抽样分布。尽管在大多数经济计量学应用中，同方差正态误差的假设似乎是不合理的，但是精确的抽样分布具有理论上的意义，根据这些分布计算的 p 值也经常出现在回归软件的输出结果中。

最后两节转向多元回归模型系数的有效估计理论的研究。16.5 节将高斯—马尔可夫定理推广到多元回归的情形。16.6 节提出了广义最小二乘法(GLS)。

必要的数学知识。本章对线性模型的处理使用了矩阵符号和线性代数的基本工具，并假定读者已修了线性代数的入门课程。附录 16.1 复习了本章中用到的向量、矩阵以及矩阵运算等相关知识。此外，16.1 节要用多元微积分的知识推导 OLS 估计量。

16.1 用矩阵符号表示的线性多元回归模型与 OLS 估计量

多元线性回归模型和 OLS 估计量都可以利用矩阵符号进行简洁地表示。

16.1.1 用矩阵符号表示的多元回归模型

总体多元回归模型(见重要概念 5.2)是:

$$Y_i = \beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki} + u_i, i = 1, \cdots, n \tag{16.1}$$

要用矩阵形式表示它,定义下列向量和矩阵:

$$\boldsymbol{Y} = \begin{pmatrix} Y_1 \\ Y_2 \\ \vdots \\ Y_n \end{pmatrix}, \boldsymbol{U} = \begin{pmatrix} u_1 \\ u_2 \\ \vdots \\ u_n \end{pmatrix}, \boldsymbol{X} = \begin{pmatrix} 1 & X_{11} & \cdots & X_{k1} \\ 1 & X_{12} & \cdots & X_{k2} \\ \vdots & \vdots & \ddots & \vdots \\ 1 & X_{1n} & \cdots & X_{kn} \end{pmatrix} = \begin{pmatrix} \boldsymbol{X}_1' \\ \boldsymbol{X}_2' \\ \vdots \\ \boldsymbol{X}_n' \end{pmatrix}, \boldsymbol{\beta} = \begin{pmatrix} \beta_0 \\ \beta_1 \\ \vdots \\ \beta_k \end{pmatrix} \tag{16.2}$$

因此,$\boldsymbol{Y}$ 为 $n \times 1$,$\boldsymbol{X}$ 为 $n \times (k+1)$,$\boldsymbol{U}$ 为 $n \times 1$,$\boldsymbol{\beta}$ 为 $(k+1) \times 1$。所有的矩阵和向量都是用黑体字表示的,在这种符号下:

■$\boldsymbol{Y}$ 是因变量的 n 个观测值的 $n \times 1$ 维向量。

■$\boldsymbol{X}$ 是 $k+1$ 个回归因子(包括截距的"常数"回归因子)的 n 个观测值的 $n \times (k+1)$ 阶矩阵。

■$(k+1) \times 1$ 维列向量 $\boldsymbol{X}_i$ 是 $k+1$ 个回归因子的第 i 个观测值,即 $\boldsymbol{X}_i' = (1 \quad X_{1i} \quad \cdots \quad X_{ki})$,其中 $\boldsymbol{X}_i'$ 表示 $\boldsymbol{X}_i$ 的转置。

■$\boldsymbol{U}$ 是 n 个误差项的 $n \times 1$ 维向量。

■$\boldsymbol{\beta}$ 是 $k+1$ 个未知回归系数的 $(k+1) \times 1$ 维向量。

对第 i 个观测值,利用向量 $\boldsymbol{\beta}$ 和 $\boldsymbol{X}_i$ 表示的多元回归模型(16.1)可写为:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, i = 1, \cdots, n \tag{16.3}$$

在公式(16.3)中,第一个回归因子是"常数"回归因子,它总是取值为 1,它的系数就是截距。因此,截距不单独出现在公式(16.3)中,而是作为系数向量 $\boldsymbol{\beta}$ 的第一个元素出现。



将方程中的全部 n 个观测值整合在一起,便得到用矩阵形式表示的多元回归模型:

$$\boldsymbol{Y} = \boldsymbol{X\beta} + \boldsymbol{U} \tag{16.4}$$

16.1.2 扩展的最小二乘假设

多元回归模型的扩展的最小二乘假设包括重要概念 5.4 中多元回归模型的四个最小二乘假设,以及另外两个关于同方差和误差服从正态分布的假设。当我们研究 OLS 估计量的有效性时,需要用到同方差假设;当我们研究 OLS 估计量和检验统计量的精确抽样分布时,需要用到正态性假设。

扩展的最小二乘假设在重要概念 16.1 中总结。

重要概念 16.1

多元回归模型中扩展的最小二乘假设

含有多个回归因子的线性回归模型是:

$$Y_i = \boldsymbol{X}_i'\boldsymbol{\beta} + u_i, i = 1, \cdots, n \tag{16.5}$$

扩展的最小二乘假设为:

1. $E(u_i|\boldsymbol{X}_i)=0$（$u_i$ 具有条件零均值）；

2. $(\boldsymbol{X}_i, Y_i), i=1,\cdots,n$ 是取自它们的联合分布的独立同分布（i. i. d.）抽样；

3. $\boldsymbol{X}_i$ 和 u_i 具有非零的、有限的四阶矩；

4. $\boldsymbol{X}$ 是列满秩的（不存在完全多重共线性）；

5. $\mathrm{var}(u_i|\boldsymbol{X}_i)=\sigma_u^2$（同方差性）；

6. 给定 $\boldsymbol{X}_i$ 条件下 u_i 的条件分布是正态的（正态误差）。

除了符号不同外，重要概念16.1中的前三个假设与重要概念5.4中的前三个假设完全相同。

重要概念5.4和重要概念16.1中的第四个假设看上去似乎有所不同，但事实上它们是一样的，它们只是在表述不存在完全多重共线性上的方式不同而已。回想一下，当一个回归因子能被表示为其他回归因子的完全线性组合时，就会出现完全多重共线性。用公式（16.2）的矩阵符号表示，完全多重共线性就是指 $\boldsymbol{X}$ 的一列是 $\boldsymbol{X}$ 的其他列的完全线性组合，但是如果事实果真如此，那么 $\boldsymbol{X}$ 不是列满秩的。因此说 $\boldsymbol{X}$ 的秩为 $k+1$，即秩等于 $\boldsymbol{X}$ 的列数，这是说这些回归因子不是完全多重共线的另外一种方式。

重要概念16.1中的第五个最小二乘假设是，误差项是同方差的；第六个假设是，给定 $\boldsymbol{X}_i$ 下 u_i 的条件分布是正态的。这两个假设与重要概念15.1中的最后两个假设相同，只是这里是针对多个回归因子来说的。

$\boldsymbol{U}$ 的均值向量和协方差阵的含义。重要概念16.1的最小二乘假设，隐含了在给定回归因子 $\boldsymbol{X}$ 的矩阵下，$\boldsymbol{U}$ 的条件分布的均值向量和协方差矩阵的简单表达式（附录16.2中给出了随机变量向量的均值向量和协方差矩阵的定义）。具体来说，重要概念16.1中的第一个和第二个假设隐含着 $E(u_i|\boldsymbol{X})=E(u_i|\boldsymbol{X}_i)=0$，且对于 $i\neq j$，有 $\mathrm{cov}(u_i,u_j|\boldsymbol{X})=E(u_iu_j|\boldsymbol{X})=E(u_iu_j|\boldsymbol{X}_i,\boldsymbol{X}_j)=E(u_i|\boldsymbol{X}_i)E(u_j|\boldsymbol{X}_j)=0$（见练习15.7）。第一个、第二个和第五个假设隐含着 $E(u_i^2|\boldsymbol{X})=E(u_i^2|\boldsymbol{X}_i)=\sigma_u^2$。合并这些结论，我们有：

$$\text{在假设 1 和假设 2 下}, E(\boldsymbol{U}|\boldsymbol{X})=\boldsymbol{0}_n \tag{16.6}$$

$$\text{在假设 1、假设 2 和假设 5 下}, E(\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{X})=\sigma_u^2\boldsymbol{I}_n \tag{16.7}$$

其中，$\boldsymbol{0}_n$ 是 n 维零向量，$\boldsymbol{I}_n$ 是 $n\times n$ 阶单位阵。

同样，重要概念16.1中的第一个、第二个、第五个和第六个假设意味着，以 $\boldsymbol{X}$ 为条件的 n 维随机向量 $\boldsymbol{U}$ 的条件分布是多元正态分布（定义见附录16.2），即：

$$\text{在假设 1、假设 2、假设 5 和假设 6 下，且给定 } \boldsymbol{X}, \boldsymbol{U} \text{ 的条件分布是 } N(\boldsymbol{0}_n, \sigma_u^2\boldsymbol{I}_n) \tag{16.8}$$



16.1.3 OLS 估计量

OLS 估计量使预测误差平方和 $\sum_{i=1}^{n}(Y_i-b_0-b_1X_{1i}-\cdots-b_kX_{ki})^2$ 最小（表达式（5.8））。OLS 估计量的公式可以通过以下步骤获得：首先，求预测误差平方和对系数向量中每个元素的导数；其次，令这些导数为零；最后，解出这些估计量 $\hat{\boldsymbol{\beta}}$。

对于 $j=0,\cdots,k$，预测误差平方和关于第 j 个回归系数 b_j 的导数是：

$$\frac{\partial}{\partial b_j}\sum_{i=1}^{n}(Y_i-b_0-b_1X_{1i}-\cdots-b_kX_{ki})^2=-2\sum_{i=1}^{n}X_{ji}(Y_i-b_0-b_1X_{1i}-\cdots-b_kX_{ki}) \tag{16.9}$$

其中，对于 $j=0$ 和所有的 i，有 $X_{0i}=1$。公式（16.9）右边的导数就是 $k+1$ 维向量 $-2\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\boldsymbol{b})$ 的第 j 个元素，其中，$\boldsymbol{b}$ 是由 $b_0,\cdots,b_k$ 组成的 $k+1$ 维向量。这样的导数共有 $k+1$ 个，每

一个对应于 b 的一个元素。联合起来,令每一个导数等于零,便得到了 $k+1$ 个方程系统,这就构成了 OLS 估计量的一阶条件,从而定义了 OLS 估计量 $\hat{\boldsymbol{\beta}}$,即 $\hat{\boldsymbol{\beta}}$ 是如下 $k+1$ 个方程系统的解:

$$\boldsymbol{X}'(\boldsymbol{Y}-\boldsymbol{X}\hat{\boldsymbol{\beta}})=\boldsymbol{0}_{k+1} \tag{16.10}$$

或者说,$\boldsymbol{X}'\boldsymbol{Y}=\boldsymbol{X}'\boldsymbol{X}\hat{\boldsymbol{\beta}}$。

解方程系统(16.10)便得到了用矩阵形式表示的 OLS 估计量 $\hat{\boldsymbol{\beta}}$:

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y} \tag{16.11}$$

其中,$(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 表示矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 的逆。

"不存在完全多重共线性"的作用。重要概念 16.1 中的第四个最小二乘假设表明,$\boldsymbol{X}$ 是列满秩的。反过来这又意味着,矩阵 $\boldsymbol{X}'\boldsymbol{X}$ 是满秩的,即 $\boldsymbol{X}'\boldsymbol{X}$ 是非奇异的。由于 $\boldsymbol{X}'\boldsymbol{X}$ 是非奇异的,因此它是可逆的。所以,不存在完全多重共线性这个假设确保了 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$ 存在,使得公式(16.10)具有惟一解,这样才能够计算 OLS 估计量的公式(16.11)。换句话说,如果 $\boldsymbol{X}$ 不满足列满秩,公式(16.10)便没有惟一解,而且 $\boldsymbol{X}'\boldsymbol{X}$ 是奇异的,因此,不能够计算 $(\boldsymbol{X}'\boldsymbol{X})^{-1}$,也不能根据公式(16.11)计算 $\hat{\boldsymbol{\beta}}$。

16.2 OLS 估计量和 t 统计量的渐近分布

如果样本容量很大,而且满足重要概念 16.1 中的前四个假设,那么 OLS 估计量具有一个渐近的联合正态分布,协方差矩阵的异方差稳健的估计量是一致的,异方差稳健的 OLSt 统计量具有渐近的标准正态分布。这些结论利用了多元正态分布(见附录 16.2)和中心极限定理的一个多元扩展式。

16.2.1 多元中心极限定理

重要概念 2.7 中的中心极限定理适用于一维的随机变量。为了推导 $\hat{\boldsymbol{\beta}}$ 各元素的联合渐近分布,我们需要推导一个适用于向量值的随机变量的多元中心极限定理。

多元中心极限定理将单变量中心极限定理扩展到一个向量值的随机变量 $\boldsymbol{W}$ 的观测值平均数,这里的 $\boldsymbol{W}$ 是 m 维向量。一个标量的中心极限定理与一个向量值的中心极限定理的区别,是方差的条件不同。在重要概念 2.7 中的标量情形下,要求方差既是非零的,又是有限的。在向量情形下,要求协方差矩阵是正定的和有限的。如果向量值随机变量 $\boldsymbol{W}$ 具有有限正定的协方差阵,那么对于任意非零的 m 维向量 $\boldsymbol{c}$,有 $0<\mathrm{var}(\boldsymbol{c}'\boldsymbol{W})<\infty$(见练习 16.3)。

重要概念 16.2 中对我们将要用到的多元中心极限定理给出了严密的陈述。

重要概念 16.2

多元中心极限定理

假设 $\boldsymbol{W}_1,\cdots,\boldsymbol{W}_n$ 是独立同分布的 m 维随机变量,其均值向量 $E(\boldsymbol{W}_i)=\boldsymbol{\mu}_W$,协方差阵 $E[(\boldsymbol{W}_i-\boldsymbol{\mu}_W)(\boldsymbol{W}_i-\boldsymbol{\mu}_W)']=\boldsymbol{\Sigma}_W$,其中,$\boldsymbol{\Sigma}_W$ 是正定且有限的。令 $\overline{\boldsymbol{W}}=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{W}_i$,那么 $\sqrt{n}(\overline{\boldsymbol{W}}-\boldsymbol{\mu}_W)\xrightarrow{d}N(\boldsymbol{0}_m,\boldsymbol{\Sigma}_W)$。

其中，$(\hat{\Sigma}_{\hat{\beta}})_{jj}$是$\hat{\Sigma}_{\hat{\beta}}$的第$(j,j)$个元素。

16.2.4 预测效应的置信区间

6.1节给出了两种计算两个或两个以上回归因子变化带来的预测效应标准误的方法。这些标准误有简洁的矩阵表达式，从而预测效应的置信区间也可用这种简洁的矩阵形式表示。

考虑回归因子的第i个观测值从某个初始值(如$\boldsymbol{X}_{i,0}$)变化到另一个新值(如$\boldsymbol{X}_{i,0}+\boldsymbol{d}$)，因此$\boldsymbol{X}_i$的变化为$\Delta\boldsymbol{X}_i=\boldsymbol{d}$，其中$\boldsymbol{d}$为$k+1$维向量。$\boldsymbol{X}$的这一变化可能涉及多个回归因子(即$\boldsymbol{X}_i$的多个元素)。例如，如果回归因子中的两个，一个是某一自变量的值，另一个是该自变量平方的值，那么$\boldsymbol{d}$就是这两个变量的初始值和变化后的值之差。

$\boldsymbol{X}_i$的这种变化的期望效应是$\boldsymbol{d}'\boldsymbol{\beta}$，这个效应的估计量是$\boldsymbol{d}'\hat{\boldsymbol{\beta}}$。由于正态分布随机变量的线性组合本身仍然服从正态分布，因此，$\sqrt{n}(\boldsymbol{d}'\hat{\boldsymbol{\beta}}-\boldsymbol{d}'\boldsymbol{\beta})=\boldsymbol{d}'\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(0,\boldsymbol{d}'\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})}\boldsymbol{d})$。所以，这个预测效应的标准误是$(\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{d})^{1/2}$。于是，这个预测效应的95%的置信区间是：

$$\boldsymbol{d}'\hat{\boldsymbol{\beta}}\pm1.96\sqrt{\boldsymbol{d}'\hat{\boldsymbol{\Sigma}}_{\hat{\boldsymbol{\beta}}}\boldsymbol{d}} \tag{16.19}$$

16.2.5 t统计量的渐近分布

重要概念5.6给出了利用公式(16.18)中的异方差稳健的标准误构造的检验零假设$\beta_j=\beta_{j,0}$的t统计量。t统计量服从渐近标准正态分布的这个争论，类似于15.3节中给出的关于单个回归因子模型的争论。

16.3 联合假设的检验

5.7节考虑了涉及多个约束条件的联合假设的检验，其中，每个约束条件涉及一个系数，而5.8节考察了涉及两个或两个以上系数的单一约束条件的检验。16.1节的矩阵设置可以将这两种类型的假设统一表示为对系数向量的线性约束，在这里，每个约束可以涉及多个系数。在重要概念16.1中的前四个最小二乘假设下，检验这些假设的异方差稳健的OLSF统计量在零假设下服从$F_{q,\infty}$渐近分布。

第5部分

16.3.1 用矩阵表示的联合假设

考虑一个联合假设：系数是线性的，并且施加q个约束条件，其中$q\leqslant k+1$。q个约束中的任何一个都可能涉及一个或多个回归系数。这个联合的零假设可以用矩阵符号写为：

$$\boldsymbol{R\beta}=\boldsymbol{r} \tag{16.20}$$

其中，$\boldsymbol{R}$是一个行满秩的$q\times(k+1)$的非随机矩阵，$\boldsymbol{r}$是一个非随机的$q\times1$向量。$\boldsymbol{R}$的行数为q，它是零假设下所施加的约束条件个数。



公式(16.20)中的零假设包含了5.7节和5.8节中所考虑的所有零假设。例如，5.7节中所考虑的一类联合零假设是$\beta_0=0,\beta_1=0,\cdots,\beta_{q-1}=0$。要将这个联合假设写成公式(16.20)的形式，只需令$\boldsymbol{R}=[\boldsymbol{I}_q\boldsymbol{0}_{q\times(k+1-q)}]$和$\boldsymbol{r}=\boldsymbol{0}_q$即可。

公式(16.20)中的这种公式化表示同样捕捉了5.8节中涉及多个回归系数的约束条件。例如，如果$k=2$，那么只需令$\boldsymbol{R}=[0\ 1\ 1]$，$\boldsymbol{r}=1$和$q=1$，假设$\beta_1+\beta_2=1$，就可以写成公

式(16.20)的形式。

16.3.2 F 统计量的渐近分布

检验公式(16.20)中的联合假设的异方差稳健的 F 统计量为:

$$F=(R\hat{\boldsymbol{\beta}}-r)'(R\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}R')^{-1}(R\hat{\boldsymbol{\beta}}-r)/q \qquad (16.21)$$

如果重要概念 16.1 中的前四个假设成立,那么在零假设下:

$$F\xrightarrow{d}F_{q,\infty} \qquad (16.22)$$

把 $\hat{\boldsymbol{\beta}}$ 的渐近正态性和协方差阵的异方差稳健的估计量 $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$ 的一致性相结合,便可得到这一结果。具体来说,首先注意公式(16.12)和附录 16.2 中的公式(16.48)意味着,在零假设下,$\sqrt{n}(R\hat{\boldsymbol{\beta}}-r)=\sqrt{n}R(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})\xrightarrow{d}N(\mathbf{0},R\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)}R')$。由公式(16.51)可得,在零假设下,$(R\hat{\boldsymbol{\beta}}-r)'(R\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}R')^{-1}(R\hat{\boldsymbol{\beta}}-r)=[\sqrt{n}R(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]'[R\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)}R']^{-1}[\sqrt{n}R(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\xrightarrow{d}\chi_q^2$,然而,因为 $\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\beta}-\beta)}\xrightarrow{p}\boldsymbol{\Sigma}_{\sqrt{n}(\hat{\beta}-\beta)}$,由 Slutsky 定理可得,$[\sqrt{n}R(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]'[R\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\beta}-\beta)}R']^{-1}[\sqrt{n}R(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})]\xrightarrow{d}\chi_q^2$,或者说(因为 $\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}=\hat{\boldsymbol{\Sigma}}_{\sqrt{n}(\hat{\beta}-\beta)}/n$),$F\xrightarrow{d}\chi_q^2/q$,它反过来仍然服从 $F_{q,\infty}$ 分布。

16.3.3 多元系数的置信集

正如 5.9 节中所讨论的,$\boldsymbol{\beta}$ 的两个或两个以上元素的渐近有效的置信集,可以构造为在零假设下不能被 F 统计量拒绝的值的集合。原则上,这样的集合可以通过对多个 $\boldsymbol{\beta}$ 值重复估计 F 统计量的值来计算。但是,和单个系数置信区间的情况一样,要获得该置信集的一个显性公式,变换该检验统计量的公式是很简单的。

下面给出构造 $\boldsymbol{\beta}$ 的两个或多个元素的置信集的程序。设 $\boldsymbol{\delta}$ 表示 q 维向量,它由我们欲对其构造置信集的回归系数所组成。例如,如果我们要对回归系数 β_1 和 β_2 构造置信集,那么 $q=2$,且 $\boldsymbol{\delta}=(\beta_1\ \ \beta_2)'$。一般地说,我们记 $\boldsymbol{\delta}=R\boldsymbol{\beta}$,其中,矩阵 R 由 0 和 1 组成(如在公式(16.20)后面所讨论的)。检验假设 $\boldsymbol{\delta}=\boldsymbol{\delta}_0$ 的 F 统计量是 $F=(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)'(R\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}R')^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta}_0)/q$,其中 $\hat{\boldsymbol{\delta}}=R\hat{\boldsymbol{\beta}}$。$\boldsymbol{\delta}$ 的一个 95% 的置信集是不能被 F 统计量拒绝的 $\boldsymbol{\delta}_0$ 值的集合,即当 $\boldsymbol{\delta}=R\boldsymbol{\beta}$ 时,$\boldsymbol{\delta}$ 的一个 95% 的置信集是:

$$\{\boldsymbol{\delta}:(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})'(R\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}R')^{-1}(\hat{\boldsymbol{\delta}}-\boldsymbol{\delta})/q\le c\} \qquad (16.23)$$

其中,c 是 $F_{q,\infty}$ 分布的第 95 个百分位数(5% 的临界值)。

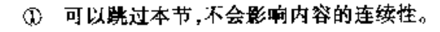
当表达式(16.23)中的不等式变为等式时,表达式(16.23)中的集合由该椭圆内所有的点构成(当 $q>2$ 时,这是个椭球)。因此,$\boldsymbol{\delta}$ 的置信集可以通过求解表达式(16.23)的椭圆边界得到。



16.4 含有正态误差的回归统计量的分布[1]

在样本容量很大时,通过大数定律和中心极限定理确立的 16.2 节和 16.3 节中给出的分布是适用的。但是,如果以 X 为条件的误差是同方差的且服从正态分布,那么以 X 为条件的 OLS 估计量服从多元正态分布。另外,该回归标准误平方的抽样分布,与自由度为 $n-$





① 可以跳过本节,不会影响内容的连续性。

$k-1$ 的卡方分布成比例，仅适用于同方差的 OLSt 统计量服从自由度为 $n-k-1$ 的学生 t 分布，而仅适用于同方差的 F 统计量服从 $F_{q,n-k-1}$ 分布。下面先给出本节讨论 OLS 回归统计量将用到的一些特殊的矩阵公式。

16.4.1 OLS 回归统计量的矩阵表示

OLS 预测值、残差以及残差平方和具有简洁的矩阵表示。这些矩阵表示利用了两个矩阵 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$。

矩阵 P_X 和 M_X。多元回归模型中 OLS 的代数运算依赖于两个对称的 $n\times n$ 阶矩阵 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$：

$$\boldsymbol{P_X}=\boldsymbol{X}(\boldsymbol{X'X})^{-1}\boldsymbol{X'} \tag{16.24}$$

$$\boldsymbol{M_X}=\boldsymbol{I_n}-\boldsymbol{P_X} \tag{16.25}$$

如果 $\boldsymbol{C}$ 是平方形式的，且 $\boldsymbol{CC}=\boldsymbol{C}$，那么矩阵 $\boldsymbol{C}$ 是幂等(idempotent)阵。由于 $\boldsymbol{P_X}=\boldsymbol{P_XP_X}$，$\boldsymbol{M_X}=\boldsymbol{M_XM_X}$（见练习 16.5），又由于 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$ 是对称的，因此，$\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$ 是对称的幂等阵。

直接根据公式(16.24)和公式(16.25)中给出的定义，矩阵 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$ 具有另外一些有用的性质：

$$\boldsymbol{P_XX}=\boldsymbol{X},\boldsymbol{M_XX}=\boldsymbol{0}_{n\times(k+1)};rank(\boldsymbol{P_X})=k+1,rank(\boldsymbol{M_X})=n-k-1 \tag{16.26}$$

其中，$rank(\boldsymbol{P_X})$ 是 $\boldsymbol{P_X}$ 的秩。

矩阵 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$ 可以将一个 n 维向量 $\boldsymbol{Z}$ 分解为两部分：一部分被 $\boldsymbol{X}$ 的列向量横穿，一部分与 $\boldsymbol{X}$ 的各列向量正交。换句话说，$\boldsymbol{P_XZ}$ 是 $\boldsymbol{Z}$ 在 $\boldsymbol{X}$ 列向量所横跨的空间上的投影，$\boldsymbol{M_XZ}$ 是 $\boldsymbol{Z}$ 与 $\boldsymbol{X}$ 各列向量正交的部分。

OLS 预测值和残差。对于 OLS 预测值和残差，矩阵 $\boldsymbol{P_X}$ 和 $\boldsymbol{M_X}$ 给出了一些简单的表达式。OLS 预测值 $\hat{\boldsymbol{Y}}=\boldsymbol{X}\hat{\boldsymbol{\beta}}$以及 OLS 残差 $\hat{\boldsymbol{U}}=\boldsymbol{Y}-\hat{\boldsymbol{Y}}$ 可被表示为（见练习 16.5）：

$$\hat{\boldsymbol{Y}}=\boldsymbol{P_XY} \tag{16.27}$$

$$\hat{\boldsymbol{U}}=\boldsymbol{M_XY}=\boldsymbol{M_XU} \tag{16.28}$$

利用公式(16.27)和公式(16.28)中的表达式可以很简单地证明，OLS 残差和预测值是正交的，即公式(4.58)成立：$\hat{\boldsymbol{Y}}'\hat{\boldsymbol{U}}=\boldsymbol{Y}'\boldsymbol{P_X}'\boldsymbol{M_XY}=0$，这里第二个等式根据 $\boldsymbol{P_X}'\boldsymbol{M_X}=\boldsymbol{0}_{n\times n}$得到，反过来它又是根据表达式(16.26)中的 $\boldsymbol{M_XX}=\boldsymbol{0}_{n\times(k+1)}$得到的。

回归的标准误。在 5.10 节中所定义的 SER 为 $s_{\hat{u}}$，这里：

$$s_{\hat{u}}^2=\frac{1}{n-k-1}\sum_{i=1}^{n}\hat{u}_i^2=\frac{1}{n-k-1}\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}=\frac{1}{n-k-1}\boldsymbol{U'M_XU} \tag{16.29}$$

其中，最后一个等式成立是由于 $\hat{\boldsymbol{U}}'\hat{\boldsymbol{U}}=(\boldsymbol{M_XU})'(\boldsymbol{M_XU})=\boldsymbol{U'M_XM_XU}=\boldsymbol{U'M_XU}$ 得到（因为 $\boldsymbol{M_X}$ 是对称且幂等的）。

16.4.2 含有正态误差的 $\hat{\boldsymbol{\beta}}$ 的分布

由于 $\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+(\boldsymbol{X'X})^{-1}\boldsymbol{X'U}$（见公式(16.14)），而且根据假设，以 $\boldsymbol{X}$ 为条件的 $\boldsymbol{U}$ 的分布为 $N(\boldsymbol{0}_n,\sigma_u^2\boldsymbol{I}_n)$（见表达式(16.8)），因此给定 $\boldsymbol{X}$ 下 $\hat{\boldsymbol{\beta}}$ 的条件分布是均值为$\boldsymbol{\beta}$ 的多元正态分布。以 $\boldsymbol{X}$ 为条件的$\hat{\boldsymbol{\beta}}$ 的协方差矩阵为 $\boldsymbol{\Sigma}_{\hat{\beta}|X}=E[(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})'|\boldsymbol{X}]=E[(\boldsymbol{X'X})^{-1}\boldsymbol{X'UU'X}(\boldsymbol{X'X})^{-1}|\boldsymbol{X}]=(\boldsymbol{X'X})^{-1}\boldsymbol{X'}(\boldsymbol{I}_n\sigma_u^2)\boldsymbol{X}(\boldsymbol{X'X})^{-1}=\sigma_u^2(\boldsymbol{X'X})^{-1}$。因此，在重要概念 16.1 中的所有六个假设下，给定 $\boldsymbol{X}$ 下 $\hat{\boldsymbol{\beta}}$ 的条件分布是：

$$\hat{\boldsymbol{\beta}} \sim N(\boldsymbol{\beta}, \boldsymbol{\Sigma}_{\hat{\beta}|X})，其中，\boldsymbol{\Sigma}_{\hat{\beta}|X} = \sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \tag{16.30}$$

16.4.3 s_u^2的分布

如果重要概念 16.1 中的所有六个假设都成立，那么 s_u^2具有精确的抽样分布，这个分布与自由度为 $n-k-1$ 的卡方分布成比例，即：

$$s_u^2 \sim \chi_{n-k-1}^2 \times \sigma_u^2/(n-k-1) \tag{16.31}$$

证明表达式(16.31)需要从公式(16.29)开始。因为 $\boldsymbol{U}$ 服从以 $\boldsymbol{X}$ 为条件的正态分布，又因为 $\boldsymbol{M}_X$ 为对称幂等阵，二次型 $\boldsymbol{U}'\boldsymbol{M}_X\boldsymbol{U}/\sigma_u^2$ 服从自由度等于 $\boldsymbol{M}_X$ 秩的卡方分布（见附录 16.2 中的公式(16.52)）。由表达式(16.26)知，$\boldsymbol{M}_X$ 的秩为 $n-k-1$。因此 $\boldsymbol{U}'\boldsymbol{M}_X\boldsymbol{U}/\sigma_u^2$ 服从精确的χ_{n-k-1}^2分布，从而表达式(16.31)成立。

自由度调整确保了 s_u^2是无偏的。一个服从χ_{n-k-1}^2分布的随机变量的期望为 $n-k-1$，因此，$E(\boldsymbol{U}'\boldsymbol{M}_X\boldsymbol{U}) = (n-k-1)\sigma_u^2$，从而，$E(s_u^2) = \sigma_u^2$。

16.4.4 仅适用于同方差的标准误

以 $\boldsymbol{X}$ 为条件的$\hat{\boldsymbol{\beta}}$ 的协方差矩阵的仅适用于同方差的估计量$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}}$，可以通过用样本方差 s_u^2 代替表达式(16.30)中的$\boldsymbol{\Sigma}_{\hat{\beta}|X}$表达式中的总体方差 σ_u^2 得到，因此：

$$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}} = s_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1} \quad （仅适用于同方差的） \tag{16.32}$$

给定 $\boldsymbol{X}$ 下 $\hat{\beta}_j$ 的正态条件分布方差的估计量是$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}}$的$(j,j)$元素，所以，$\hat{\beta}_j$ 的仅适用于同方差的标准误是$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}}$的第 j 个对角元素的平方根，即$\hat{\beta}_j$ 的仅适用于同方差的标准误是：

$$\widetilde{SE}(\hat{\beta}_j) = \sqrt{(\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}})_{jj}} \quad （仅适用于同方差的） \tag{16.33}$$

16.4.5 t 统计量的分布

设 $\tilde{t}$ 为检验假设$\beta_j = \beta_{j,0}$的 t 统计量，它是利用仅适用于同方差的标准误构造的，即设：

$$\tilde{t} = \frac{\hat{\beta}_j - \beta_{j,0}}{\sqrt{(\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}})_{jj}}} \tag{16.34}$$

在重要概念 16.1 中的全部六个扩展的最小二乘假设下，$\tilde{t}$ 的精确抽样分布是自由度为 $n-k-1$ 的学生 t 分布，即：

$$\tilde{t} \sim t_{n-k-1} \tag{16.35}$$

对表达式(16.35)的证明在附录 16.4 中给出。

16.4.6 F 统计量的分布

如果重要概念 16.1 中的全部六个最小二乘假设都成立，那么利用协方差矩阵的仅适用于同方差的估计量所构造的检验公式(16.20)中假设的 F 统计量，在零假设下服从精确的$F_{q,n-k-1}$分布。

仅适用于同方差的 F 统计量。仅适用于同方差的 F 统计量类似于表达式(16.21)中异方差稳健的 F 统计量，不同的是，这里使用了仅适用于同方差的估计量$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}}$，而不是异方差稳健的估计量$\hat{\boldsymbol{\Sigma}}_{\hat{\beta}}$。将表达式$\tilde{\boldsymbol{\Sigma}}_{\hat{\beta}} = s_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$代入表达式(16.21)中 F 统计量的表达式，便得



到检验公式(16.20)中零假设的仅适用于同方差的 F 统计量。

$$\tilde{F}=\frac{(R\hat{\beta}-r)'[R(X'X)^{-1}R']^{-1}(R\hat{\beta}-r)/q}{s_{\hat{u}}^2} \tag{16.36}$$

如果重要概念 16.1 中的全部六个假设都成立,那么在零假设下:

$$\tilde{F}\sim F_{q,n-k-1} \tag{16.37}$$

表达式(16.37)的证明在附录 16.4 中给出。

经验规则的 F 统计量公式。公式(16.36)中的 F 统计量被称为 Wald 形式的 F 统计量(以统计学家 Abraham Wald 的名字命名)。尽管附录 5.3 中给出的经验规则的 F 统计量公式似乎与公式(16.36)中的 Wald 统计量公式很不相同,但实际上经验规则 F 统计量和 Wald 统计量只是同一个统计量的两种形式。也就是说,这两个表达式是等价的,进一步的讨论见 Greene(1997,第 7 章)。

16.5 含有同方差误差的 OLS 估计量的有效性

在多元回归的高斯—马尔可夫条件下,β 的 OLS 估计量在所有线性条件无偏估计量中是有效的,也就是说,OLS 估计量是 BLUE。

16.5.1 多元回归的高斯—马尔可夫条件

多元回归的高斯—马尔可夫条件(Gauss - Markov conditions for multiple regression)是:

$$\begin{aligned}&(\text{i})E(U|X)=0_n\\&(\text{ii})E(UU'|X)=\sigma_u^2I_n\\&(\text{iii})X\text{ 是列满秩的}\end{aligned} \tag{16.38}$$

多元回归的高斯—马尔可夫条件反过来被重要概念 16.1(见表达式(16.6)和表达式(16.7))中的前五个假设所隐含。表达式(16.8)中的条件将高斯—马尔可夫条件由单个回归因子情形推广到多元情形(通过利用矩阵符号,将表达式(15.24)中的第二个和第三个高斯—马尔可夫条件合并成表达式(16.38)中的一个条件(ii))。

16.5.2 线性条件无偏估计量

我们先描述线性无偏估计量的类别,然后证明 OLS 属于此类。

线性条件无偏估计量的类别。如果 β 的估计量是 $Y_1,\cdots,Y_n$ 的线性函数,则称该估计量是线性的。因此,如果估计量 $\tilde{\beta}$ 可以写成如下形式:

$$\tilde{\beta}=A'Y \tag{16.39}$$

那么它与 Y 是线性的关系。这里的 A 是一个可能依赖于 X 和非随机常数但不依赖于 Y 的 $n\times(k+1)$ 阶的权数矩阵。如果给定 X,一个估计量的条件抽样分布的均值为 β,则该估计量是条件无偏的。也就是说,如果 $E(\tilde{\beta}|X)=\beta$,那么 $\tilde{\beta}$ 是条件无偏的。

OLS 估计量是线性的和条件无偏的。比较公式(16.11)和公式(16.39)可知,该 OLS 估计量与 Y 是线性的关系,具体地说,$\hat{\beta}=\hat{A}'Y$,其中 $\hat{A}=X(X'X)^{-1}$。为了证明 $\hat{\beta}$ 是条件无偏的,回想一下公式(16.14)中 $\hat{\beta}=\beta+(X'X)^{-1}X'U$。对这个表达式的两边同时取条件期望得 $E(\hat{\beta}|X)=\beta+E[(X'X)^{-1}X'U|X]=\beta+(X'X)^{-1}X'E(U|X)=\beta$,其中,最后一个等式根据第一个高斯—马尔可夫条件 $E(U|X)=0$ 得到。



16.5.3 多元回归的高斯—马尔可夫定理

多元回归的高斯—马尔可夫定理（The Gauss - Markov theorem for multiple regression）给出了 OLS 估计量在线性条件无偏估计量中是有效的所需满足的条件。但是，因为$\hat{\boldsymbol{\beta}}$是个向量，而它的“方差”是个协方差矩阵，所以，这里存在一个细微的问题。当一个估计量的“方差”是个矩阵时，认为一个估计量比另一个估计量的方差小，其含义是什么呢？

高斯—马尔可夫定理可以回答这个问题，它是通过比较$\boldsymbol{\beta}$各元素一个线性组合的候选估计量的方差与相应的$\hat{\boldsymbol{\beta}}$线性组合的方差来完成的。具体来说，设$\boldsymbol{c}$是个$k+1$维向量，考虑这样一个问题：一方面用候选的估计量$\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$（$\tilde{\boldsymbol{\beta}}$是一个线性条件无偏估计量）估计线性组合$\boldsymbol{c}'\boldsymbol{\beta}$，另一方面用$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$估计$\boldsymbol{c}'\boldsymbol{\beta}$。由于$\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$和$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$都是标量，同时都是$\boldsymbol{c}'\boldsymbol{\beta}$的线性条件无偏估计量，因此，现在比较它们的方差就有意义了。

多元回归的高斯—马尔可夫定理表明，$\boldsymbol{c}'\boldsymbol{\beta}$的OLS估计量是有效的，即OLS估计量$\boldsymbol{c}'\hat{\boldsymbol{\beta}}$在所有线性条件无偏估计量$\boldsymbol{c}'\tilde{\boldsymbol{\beta}}$中具有最小的条件方差。值得注意的是，不管线性组合的形式是什么，上述命题都成立。正是在这个意义下，多元回归中的 OLS 估计量是 BLUE。

高斯—马尔可夫定理的陈述在重要概念 16.3 中给出，并在附录 16.5 中进行了证明。

重要概念 16.3

多元回归的高斯—马尔可夫定理

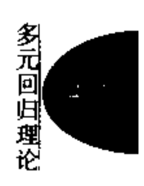
假设表达式（16.38）中多元回归的高斯—马尔可夫条件都成立，那么OLS估计量$\hat{\boldsymbol{\beta}}$是BLUE。也就是说，设$\tilde{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的一个线性条件无偏估计量，$\boldsymbol{c}$是一个非随机的$k+1$维向量，那么对于任一非零向量$\boldsymbol{c}$，便有$\mathrm{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}}|\boldsymbol{X}) \leqslant \mathrm{var}(\boldsymbol{c}'\tilde{\boldsymbol{\beta}}|\boldsymbol{X})$。其中，对于所有的$\boldsymbol{c}$来说，当且仅当$\tilde{\boldsymbol{\beta}}=\hat{\boldsymbol{\beta}}$时，不等式中的等式才成立。

16.6 广义最小二乘法[①]

第5部分

独立同分布的抽样假定适用于许多实际问题。例如，假设Y_i和$\boldsymbol{X}_i$对应于个体的信息，比如说人们的收入、受教育水平和个人特征等，这里的个体是通过简单随机抽样从总体中抽取的。在这种情形下，由于这个简单随机抽样方案，$(\boldsymbol{X}_i, Y_i)$必然是独立同分布的。因为对于$i \neq j$，$(\boldsymbol{X}_i, Y_i)$和$(\boldsymbol{X}_j, Y_j)$是独立分布的，所以，u_i和u_j也是独立分布的。反过来这又意味着，对于$i \neq j$，u_i和u_j是不相关的。在高斯—马尔可夫假设的内容中，如果数据是以一种使观测值是独立分布的方式搜集的，那么假设$E(\boldsymbol{UU}'|\boldsymbol{X})$是对角的这个结论是适当的。

然而，经济计量学中遇到的许多抽样方案并不能保证一定会得到独立的观测值，相反会引致误差项在不同的观测值之间是相关的。最为突出的例子是当数据是对同一个实体在不同时间被抽取时，即当数据是时间序列数据时，就会发生上述结果。正如 13.3 节中所讨论的，在涉及时间序列数据的回归中，许多被忽略的因素在前后期之间是相关的，这就可能导致回归误差项（它代表了那些被忽略的因素）在前后期观测值之间是相关的。换句话说，某



① 在 13.5 节的分布滞后时间序列回归中介绍了 GLS 估计量。这里的这个介绍是独立的 GLS 的数学处理，可独立于 13.5 节来学习，但先学习那一节有助于使这些思想变得更清楚。

一期的误差项与下一期的误差项一般不会是独立分布的;相反,一个时期的误差项可能与下一期的误差项相关。

相关误差项的存在为基于 OLS 的统计推断引来了两个问题。首先,由 OLS 所产生的异方差稳健的标准误和仅适用于同方差的标准误都不能为统计推断提供一个有效的基础。解决这个问题的方法是,使用对观测值之间误差项的异方差和相关系数同时稳健的标准误。这个问题——异方差—自相关——致性(HAC)协方差矩阵估计——是 13.4 节的主题,这里我们不打算做进一步探讨了。

其次,如果误差项在观测值之间是相关的,那么 $E(\boldsymbol{UU}'|\boldsymbol{X})$ 不是对角的,表达式(16.38)中的第二个高斯—马尔可夫条件不成立,从而 OLS 不是 BLUE。在本节,我们研究一个估计量——广义最小二乘(generalized least squares,简写为 GLS),当误差项的条件协方差矩阵不再与单位阵成比例时,这个 GLS 是 BLUE(至少在渐近意义下)。GLS 的一个特殊情形是 15.5 节中所讨论的加权最小二乘,在那里,条件协方差矩阵是对角的,且第 i 个对角元素是 $\boldsymbol{X}_i$ 的函数。和 WLS 一样,GLS 对回归模型进行了变换,这使得变换后的模型的误差项满足高斯—马尔可夫条件。这个 GLS 估计量就是变换后的模型中系数的 OLS 估计量。

16.6.1 GLS 假设

要使 GLS 是有效的,需要有四个假设条件。第一个 GLS 假设是,以 $\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n$ 为条件的 u_i 具有零均值,即:

$$E(\boldsymbol{U}|\boldsymbol{X})=\boldsymbol{0}_n \tag{16.40}$$

这个假设隐含在重要概念 16.1 中的前两个最小二乘假设中,也就是说,如果 $E(u_i|\boldsymbol{X}_i)=0$ 且 $(\boldsymbol{X}_i,Y_i)$,$i=1,\cdots,n$ 是独立同分布的,那么 $E(\boldsymbol{U}|\boldsymbol{X})=\boldsymbol{0}_n$。可是,在 GLS 中,我们并不要求保持这个独立同分布假设,因为 GLS 的一个目的就是处理观测值之间误差项的相关性问题。介绍完 GLS 估计量之后,我们将讨论公式(16.40)中那个假设的意义。

第二个 GLS 假设是,给定 $\boldsymbol{X}$ 条件下,$\boldsymbol{U}$ 的条件协方差矩阵是 $\boldsymbol{X}$ 的某种函数。

$$E(\boldsymbol{UU}'|\boldsymbol{X})=\boldsymbol{\Omega}(\boldsymbol{X}) \tag{16.41}$$

其中,$\boldsymbol{\Omega}(\boldsymbol{X})$ 是 $\boldsymbol{X}$ 的 $n\times n$ 阶正定矩阵值函数。

与这个假设相关联的 GLS 有两个主要应用。第一个应用是带有异方差误差的独立抽样,其中 $\boldsymbol{\Omega}(\boldsymbol{X})$ 是个对角阵,对角元素是 $\lambda h(\boldsymbol{X}_i)$,这里的 λ 是个常数,h 是一个函数。15.5 节中讨论过这种情形,结论是 GLS 是 WLS。

第二个应用是针对序列相关的同方差误差项。实际上,在这种情形下,人们提出了一种处理序列相关的模型,例如,误差项只与它相邻的误差项相关的模型,所以,$corr(u_i,u_{i-1})=\rho\neq0$,但是,如果 $|i-j|\geq2$,那么 $corr(u_i,u_{i-1})=0$。在这种情形下,$\boldsymbol{\Omega}(\boldsymbol{X})$ 对角线元素为 σ_u^2,第一个次对角线元素为 $\rho\sigma_u^2$,其他元素均为 0。因此,$\boldsymbol{\Omega}(\boldsymbol{X})$ 并不依赖于 $\boldsymbol{X}$,$\boldsymbol{\Omega}_{ii}=\sigma_u^2$;对于 $|i-j|=1$,$\boldsymbol{\Omega}_{ij}=\rho\sigma_u^2$;对于 $|i-j|>1$,$\boldsymbol{\Omega}_{ij}=0$。序列相关的其他模型,包括一阶自回归模型,在 13.5 节的 GLS 内容中进行了进一步讨论(也可见练习 16.8)。

在前面出现的关于截面数据最小二乘假设所有假设中的一个假设是,$\boldsymbol{X}_i$ 和 u_i 具有非零的有限的四阶矩。在 GLS 情形中,具体的矩假设需要根据函数 $\boldsymbol{\Omega}(\boldsymbol{X})$ 的性质来证明渐近结果。特定的矩假设还依赖于我们是研究 GLS 估计量还是考察 t 或 F 统计量,而且对矩的要求还依赖于 $\boldsymbol{\Omega}(\boldsymbol{X})$ 是否是已知的,或者说是否已估计出了它的参数。因为这些假设随着具体情况和具体模型不同而不同,所以,在此我们不打算给出具体的矩假设。关于 GLS 大样

本性质的讨论,假设这样的矩条件对解决我们目前的问题是适合的。为完整起见,作为 GLS 的第三个假设,简单地假设 X_t 和 u_t 满足合适的矩条件。

GLS 的第四个假设是,X 是列满秩的,即回归因子不是完全多重共线性的。

GLS 的假设被概括在重要概念 16.4 中。

重要概念 16.4

GLS 假设

在线性回归模型 $Y = X\beta + U$ 中,相应的 GLS 假设是:

1. $E(U|X) = 0_n$;
2. $E(UU'|X) = \Omega(X)$,其中 $\Omega(X)$ 是可能依赖于 X 的 $n \times n$ 正定矩阵;
3. X_t 和 u_t 满足合适的矩条件;
4. X 是列满秩的(不存在完全多重共线性)。

我们分两种情况分析 GLS 估计。第一种情况,$\Omega(X)$ 是已知的。第二种情况,$\Omega(X)$ 的函数形式是已知的,其中的参数可以估计。为了简化符号,我们将函数 $\Omega(X)$ 简写为矩阵 Ω,这样 Ω 对 X 的依赖关系没有被明确地给出来。

16.6.2 Ω 已知时的 GLS

当 Ω 已知时,GLS 估计量利用 Ω 将回归模型变换成一个误差项满足高斯—马尔可夫条件的模型。具体地说,设 F 为 Ω^{-1} 的矩阵平方根,即设 F 是个满足 $F'F = \Omega^{-1}$ 的矩阵(这样的矩阵总是存在的)。F 的一个性质是 $F'\Omega F = I_n$。现在公式(16.4)的两边同时前乘 F,得到:

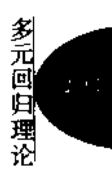
$$\widetilde{Y} = \widetilde{X}\beta + \widetilde{U} \tag{16.42}$$

其中,$\widetilde{Y} = FY$,$\widetilde{X} = FX$ 和 $\widetilde{U} = FU$。

GLS 的关键思想是,在四个 GLS 假设条件下,高斯—马尔可夫假设对公式(16.42)中变换后的回归成立。也就是说,通过用 Ω 的矩阵平方根的倒数变换所有的变量,在变换后的回归模型中,回归误差的条件均值为零,协方差矩阵等于单位阵。要在数学上证明这一点,首先注意,根据第一个GLS假设(见公式(16.40)),$E(\widetilde{U}|\widetilde{X}) = E(FU|FX) = FE(U|FX) = 0_n$。另外,$E(\widetilde{U}\widetilde{U}'|\widetilde{X}) = E[(FU)(FU)'|FX] = FE(UU'|FX)F' = F\Omega F' = I_n$,其中,第二个等式可由 $(FU)' = U'F'$ 得到,而最后一个等式由 F 的定义得出。由此推断,公式(16.42)中变换后的回归模型满足重要概念 16.3 中的高斯—马尔可夫条件。

GLS 估计量 $\hat{\beta}^{GLS}$ 就是公式(16.42)中 β 的 OLS 估计量,即 $\hat{\beta}^{GLS} = (\widetilde{X}'\widetilde{X})^{-1}(\widetilde{X}'\widetilde{Y})$。由于转换后的回归模型满足高斯—马尔可夫条件,因此,这个GLS估计量是关于 $\widetilde{Y}$ 线性的最佳条件无偏估计量。但是因为 $\widetilde{Y} = FY$,并且假设 F(在这里)是已知的,又因为 F 是可逆的(因为 Ω 是正定的),所以,关于 $\widetilde{Y}$ 是线性的那些估计量的类别与关于 Y 是线性的那些估计量的类别相同。因此,公式(16.42)中 β 的这个 OLS 估计量也是在关于 Y 是线性的那些估计量类别中的最佳条件无偏估计量。换句话说,在 GLS 假设下,GLS 估计量是 BLUE。

GLS 估计量可以直接用 Ω 表示,因此,原则上不需要计算平方根矩阵 F。因为 $\widetilde{X} = FX$,且 $\widetilde{Y} = FY$,所以,$\hat{\beta}^{GLS} = (X'F'FX)^{-1}(X'F'FY)$。但是,$F'F = \Omega^{-1}$,所以:



$$\hat{\boldsymbol{\beta}}^{GLS}=(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{Y}) \tag{16.43}$$

实际上，$\boldsymbol{\Omega}$ 一般是未知的，公式(16.43)中的 GLS 估计量一般是不可计算的，所以有时也称其为不可行的 GLS(infeasible GLS)估计量。可是，如果 $\boldsymbol{\Omega}$ 的函数形式已知，但该函数的参数却是未知的，那么 $\boldsymbol{\Omega}$ 能被估计出来，所以一个可行的 GLS 估计量可以计算出来。

16.6.3 $\boldsymbol{\Omega}$ 包含未知参数时的 GLS

如果 $\boldsymbol{\Omega}$ 是一个已知的函数，它的一些参数可以被估计出来，那么这些估计的参数可用来计算协方差矩阵 $\boldsymbol{\Omega}$ 的估计量。例如，考虑公式(16.41)后面讨论的那个时间序列的应用，其中，$\boldsymbol{\Omega}(\boldsymbol{X})$ 并不依赖于 $\boldsymbol{X}$，$\boldsymbol{\Omega}_{ii}=\sigma_u^2$，对 $|i-j|=1$，$\boldsymbol{\Omega}_{ij}=\rho\sigma_u^2$，且对 $|i-j|>1$，$\boldsymbol{\Omega}_{ij}=0$。那么 $\boldsymbol{\Omega}$ 具有两个未知参数，σ_u^2 和 ρ。这些参数可以使用残差利用先前的 OLS 回归进行估计。具体来说，可以用 $s_{\hat{u}}^2$ 估计 σ_u^2，可以用相邻的 OLS 残差对之间的样本相关系数估计 ρ。这些所估计的参数反过来又能被用来计算 $\boldsymbol{\Omega}$ 的估计量 $\hat{\boldsymbol{\Omega}}$。

一般地说，假设已有了 $\boldsymbol{\Omega}$ 的估计量 $\hat{\boldsymbol{\Omega}}$，那么基于 $\hat{\boldsymbol{\Omega}}$ 的 GLS 估计量为：

$$\hat{\boldsymbol{\beta}}^{GLS}=(\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{X})^{-1}(\boldsymbol{X}'\hat{\boldsymbol{\Omega}}^{-1}\boldsymbol{Y}) \tag{16.44}$$

公式(16.44)中的 GLS 估计量有时又被称为可行的 GLS(feasible GLS)估计量。因为如果协方差矩阵中包含的一些未知参数可以被估计的话，那么它就可以被计算出来。

16.6.4 条件零均值假设与 GLS

为了确保 OLS 估计量是一致的，第一个最小二乘假设必须满足，即 $E(u_i|\boldsymbol{X}_i)$ 必须等于零。相比之下，第一个 GLS 假设是 $E(u_i|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)=0$。换句话说，第一个 OLS 假设是，给定第 i 个回归因子的观测值，误差项具有条件零均值，而第一个 GLS 假设是，给定回归因子的所有观测值，u_i 具有条件零均值。

如 16.1 节中所讨论的，假设 $E(u_i|\boldsymbol{X}_i)=0$ 和抽样是独立同分布的假设合在一起，这意味着 $E(u_i|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)=0$。因此，如果抽样是独立同分布的，那么 GLS 就是 WLS，第一个 GLS 假设就蕴涵于重要概念 16.1 中的第一个最小二乘假设中。

然而，如果抽样不是独立同分布的，那么第一个 GLS 假设就不会由假设 $E(u_i|\boldsymbol{X}_i)=0$ 所隐含，即第一个 GLS 假设是更强的。尽管这两个条件的区别看起来似乎是微乎其微的，但是在时间序列的应用中，这是非常重要的。在 13.5 节回归因子是“当前和过去”外生的还是“严”外生的内容中，我们曾讨论了这个区别，$E(u_i|\boldsymbol{X}_1,\cdots,\boldsymbol{X}_n)=0$ 这个假设对应于严外生性。现在，我们利用矩阵符号在更一般的层次上讨论这个区别。为此，我们集中考虑 $\boldsymbol{U}$ 是同方差、$\boldsymbol{\Omega}$ 已知且 $\boldsymbol{\Omega}$ 具有非零的非对角线元素的情形。

第一个 GLS 假设的作用。要理解这些假设之间区别的来源，比较 GLS 和 OLS 的一致性讨论是有用的。

我们首先给出公式(16.43)中关于 GLS 估计量一致性问题的讨论。将公式(16.4)代入公式(16.43)，我们便有 $\hat{\boldsymbol{\beta}}^{GLS}=\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}/n)^{-1}(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}/n)$。在第一个 GLS 假设下，$E(\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U})=E[\boldsymbol{X}'\boldsymbol{\Omega}^{-1}E(\boldsymbol{U}|\boldsymbol{X})]=0_n$。此外，如果 $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}/n$ 的方差又趋向于零，且 $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{X}/n\xrightarrow{p}\tilde{\boldsymbol{Q}}$，其中 $\tilde{\boldsymbol{Q}}$ 是某种可逆矩阵，那么 $\hat{\boldsymbol{\beta}}^{GLS}\xrightarrow{p}\boldsymbol{\beta}$。非常关键地，当 $\boldsymbol{\Omega}$ 具有非对角线元素时，对于不同的 i,j，$\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}=\sum_{i=1}^{n}\sum_{j=1}^{n}\boldsymbol{X}_i(\boldsymbol{\Omega}^{-1})_{ij}u_j$ 中含有 $\boldsymbol{X}_i$ 和 u_j 的乘积，其中，$(\boldsymbol{\Omega}^{-1})_{ij}$ 表示 $\boldsymbol{\Omega}^{-1}$ 的 (i,j) 元素。因此，要使 $\boldsymbol{X}'\boldsymbol{\Omega}^{-1}\boldsymbol{U}$ 具有零均值，仅有 $E(u_i|\boldsymbol{X}_i)=0$ 是不够



的，对于所有对应于$(\boldsymbol{\Omega}^{-1})_{ij}$的非零值的$i,j$对元素，必须有$E(u_i|X_j)=0$。根据误差的协方差结构，只有$(\boldsymbol{\Omega}^{-1})_{ij}$的某些或所有的元素可能是非零的。例如，如果$u_i$服从一阶自回归（如13.5节中所讨论的），仅有的非零元素$(\boldsymbol{\Omega}^{-1})_{ij}$是那些当$|i-j|\leq 1$时的元素。可是一般地，$\boldsymbol{\Omega}^{-1}$的所有元素都可能是非零的，所以一般来说，要使$\boldsymbol{X}'\boldsymbol{\Omega}^{-1}U/n \xrightarrow{p} \boldsymbol{0}_{(k+1)\times 1}$（从而使$\hat{\boldsymbol{\beta}}^{GLS}$是一致的），我们需要$E(\boldsymbol{U}|\boldsymbol{X})=\boldsymbol{0}_n$，即第一个GLS假设必须成立。

相反，回想一下关于OLS估计量一致性的讨论。将方程(16.14)重写为$\hat{\boldsymbol{\beta}}=\boldsymbol{\beta}+(\boldsymbol{X}'\boldsymbol{X}/n)^{-1}\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i u_i$。如果$E(u_i|\boldsymbol{X}_i)=0$，那么$\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i u_i$项具有零均值，并且，如果该项的方差趋向于零，那么它依概率收敛于零。如果又有$\boldsymbol{X}'\boldsymbol{X}/n \xrightarrow{p} \boldsymbol{Q}_X$，那么$\hat{\boldsymbol{\beta}} \xrightarrow{p} \boldsymbol{\beta}$。

第一个GLS假设是有约束性的吗？第一个GLS假设要求第i个观测值的误差与所有其他观测值的回归因子不相关。这一假设在有些时间序列应用中可能是值得怀疑的。这个问题在13.6节以一个实证例子进行了讨论，即浓缩橙汁期货交割价格的变化与佛罗里达州的天气之间的关系。正如那部分所解释的，价格变化对天气回归中的误差项，与天气的当前值和过去值合理地不相关，所以第一个OLS假设成立。然而，这个误差项却可能与天气的未来值相关，从而第一个GLS假设不成立。

这个例子说明了经济时间序列中的一个普遍现象，当一个变量的当前值部分地以未来的预期值为基础设定时，就会出现这种现象。那些对未来的期望往往意味着，今天的误差项依赖于对回归因子明天的预期，而反过来明天的这种预期与回归因子明天的实际值相关。出于这种原因，第一个GLS假设实际上比第一个OLS假设强得多。因此，在某些经济时间序列数据的应用中，GLS估计量不是一致的估计量，即使OLS估计量是一致的估计量。

总结

1. 用矩阵形式表示的线性多元回归模型是$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{U}$，其中，$\boldsymbol{Y}$是因变量观测值的$n\times 1$向量，$\boldsymbol{X}$是由$k+1$个回归因子（包括常数）的$n$个观测值组成的$n\times(k+1)$矩阵，$\boldsymbol{\beta}$是未知参数的$k+1$维向量，$\boldsymbol{U}$是误差项的$n\times 1$维向量。

2. OLS估计量是$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{Y}$。在重要概念16.1中的前四个最小二乘假设下，$\hat{\boldsymbol{\beta}}$是一致的，且服从渐近正态分布。此外，如果误差项是同方差的，那么$\hat{\boldsymbol{\beta}}$的条件方差是$\text{var}(\hat{\boldsymbol{\beta}}|\boldsymbol{X})=\sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}$。

3. 对$\boldsymbol{\beta}$的一般线性约束可以写成q个方程，即$\boldsymbol{R\beta}=\boldsymbol{r}$，并且这个公式可用来检验涉及多个系数的联合假设或构造$\boldsymbol{\beta}$中各元素的置信集。

4. 当以$\boldsymbol{X}$为条件的回归误差是独立同分布的，且服从正态分布时，$\boldsymbol{\beta}$服从精确的正态分布，且仅适用于同方差的t统计量和F统计量分别服从精确的t_{n-k-1}和$F_{q,n-k-1}$分布。

5. 高斯—马尔可夫定理指出，如果误差项是同方差的，并在观测值之间条件不相关，而且如果$E(u_i|\boldsymbol{X})=0$，那么OLS估计量在线性条件无偏估计量中是有效的（OLS是BLUE）。

6. 如果误差的协方差矩阵$\boldsymbol{\Omega}$与单位阵不成比例，并且$\boldsymbol{\Omega}$是已知的，或是可以估计的，那么GLS估计量比OLS估计量渐近地有效。但是一般地说，GLS要求u_i与所有回归因子的观测值不相关，并且不像OLS那样只要求与$\boldsymbol{X}_i$不相关，在实际应用中，这个假设往往必须进行谨慎评价。



重要术语

幂等的　多元回归的高斯—马尔可夫条件　多元回归的高斯—马尔可夫定理　均值向量　广义最小二乘(GLS)　不可行的 GLS　可行的 GLS　协方差矩阵

复习概念

16.1　一位研究人员研究一组工人的收入与性别之间的关系，将回归模型设定为 $Y_i = \beta_0 + X_{1i}\beta_1 + X_{2i}\beta_2 + u_i$，其中，$X_{1i}$是个二元变量，如果第 i 个人是女性，那么它等于 1；X_{2i}也是个二元变量，如果第 i 个人是男性，那么它等于 1。对于一个 $n=5$ 的假设的观测值集合，请用表达式(16.2)的矩阵形式表示这个模型。证明 $\boldsymbol{X}$ 列向量是线性相关的，因此 $\boldsymbol{X}$ 不是列满秩的。请解释你将如何重新设定这个模型，以剔除完全多重共线性。

16.2　你在分析一个拥有 500 个观测值和一个回归因子的线性回归模型。请解释你将如何构造 β_1 的置信区间，如果：

a. 重要概念 16.1 中的假设 1 ~4 成立，但你认为假设 5 和假设 6 可能是不成立的。

b. 重要概念 16.1 中的假设 1 ~5 成立，而你认为假设 6 可能不成立(给出构造置信区间的两种方式)。

c. 假设 1 ~6 都成立。

16.3　假定重要概念 16.1 中的假设 1 ~5 都成立，但假设 6 不成立。表达式(16.31)中的结果会成立吗？请解释说明。

16.4　如果公式(16.41)成立，但不知道 $\boldsymbol{\Omega}$，你能计算出 $\boldsymbol{\beta}$ 的 BLUE 估计量吗？如果 $\boldsymbol{\Omega}$ 已知，你又怎样计算呢？

16.5　构造一个回归模型的例子，它满足假设 $E(u_i|\boldsymbol{X}_i)=0$，但 $E(\boldsymbol{U}|\boldsymbol{X})\neq\boldsymbol{0}_n$。

练习

*16.1　考虑公式(6.1)中考试成绩对收入和收入平方的总体回归。

a. 将公式(6.1)中的回归写成公式(16.4)中的矩阵形式，定义 $\boldsymbol{Y},\boldsymbol{X},\boldsymbol{U}$ 和 $\boldsymbol{\beta}$。

b. 解释如何在考试成绩与收入之间的关系是二次的备择假设下检验考试成绩与收入的关系是线性的零假设。用公式(16.20)的形式写出零假设。$\boldsymbol{R}$,$\boldsymbol{r}$ 和 q 分别是什么？

16.2　设有一个 $n=20$ 的居民户样本，一个因变量和两个回归因子的样本均值和样本协方差数据见表 16—1。

表 16—1　**样本均值和样本协方差数据表**

	样本均值	样本协方差		
		Y	X_1	X_2
$\boldsymbol{Y}$	6.39	0.26	0.22	0.32
$\boldsymbol{X_1}$	7.24		0.80	0.28
$\boldsymbol{X_2}$	4.00			2.40

a. 计算 β_0,β_1 和 β_2 的 OLS 估计值，计算 s_u^2 以及回归的 R^2。



b. 假设重要概念 16.1 中的所有六个假设都成立，在 5% 的显著性水平下检验假设 $\beta_1=0$。

16.3 假设 $\boldsymbol{W}$ 是协方差矩阵为 $\boldsymbol{\Sigma}_W$ 的 $m\times1$ 向量，其中，$\boldsymbol{\Sigma}_W$ 是有限的且是正定的。假设 $\boldsymbol{c}$ 是一个非随机的 $m\times1$ 向量，并且设 $Q=\boldsymbol{c}'\boldsymbol{W}$。

a. 证明：$\mathrm{var}(Q)=\boldsymbol{c}'\boldsymbol{\Sigma}_W\boldsymbol{c}$。

b. 设 $\boldsymbol{c}\neq\boldsymbol{0}_m$，证明 $0<\mathrm{var}(Q)<\infty$。

16.4 考虑第 4 章的回归模型：$Y_i=\beta_0+\beta_1X_i+u_i$，并假设重要概念 4.3 中的所有假设都成立。

a. 以表达式(16.2)和表达式(16.4)中给出的矩阵形式表示这个模型。

b. 证明：重要概念 16.1 中的假设 1 ~ 4 都成立。

c. 使用公式(16.11)中 $\hat{\boldsymbol{\beta}}$ 的一般公式推导重要概念 4.2 中给出的 $\hat{\beta}_0$ 和 $\hat{\beta}_1$ 的表达式。

d. 证明：表达式(16.13)中的 $\boldsymbol{\Sigma}_{\hat{\boldsymbol{\beta}}}$ 的(1,1)元素等于重要概念 4.4 中给出的 $\sigma^2_{\hat{\beta}_0}$ 的表达式。

16.5 设 $\boldsymbol{P}_X$ 和 $\boldsymbol{M}_X$ 如公式(16.24)和公式(16.25)所下的定义。

a. 证明：$\boldsymbol{P}_X\boldsymbol{M}_X=\boldsymbol{0}_{n\times n}$，且 $\boldsymbol{P}_X$ 和 $\boldsymbol{M}_X$ 是幂等的。

b. 推导公式(16.27)和公式(16.28)。

*16.6 考虑下列矩阵形式的回归模型：$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{Wy}+\boldsymbol{u}$，其中，$\boldsymbol{X}$ 为回归因子的 $n\times k_1$ 阶矩阵，$\boldsymbol{W}$ 是回归因子的一个 $n\times k_2$ 阶矩阵，那么 OLS 估计量 $\hat{\boldsymbol{\beta}}$ 可表示为：

$$\hat{\boldsymbol{\beta}}=(\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X})^{-1}(\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{Y}) \tag{16.45}$$

现在设 $\hat{\beta}_1^{BV}$ 是利用 OLS 通过估计公式(8.11)计算出来的“二元变量”固定效应估计量，并设 $\hat{\beta}_1^{DM}$ 是利用 OLS 通过估计公式(8.14)计算出来的“没有意义的”(de-meaning)固定效应估计量，其中，特定实体的样本均值已经从各自的 $\boldsymbol{X}$ 和 $\boldsymbol{Y}$ 中减去了。利用公式(16.45)证明：$\hat{\beta}_1^{BV}=\hat{\beta}_1^{DM}$。(提示：利用一个固定效应的满集，即 $D1_i,D2_i,\cdots,Dn_i$ 且不包括常数项，重新写出公式(8.11)。在 $\boldsymbol{W}$ 中包含所有的固定效应。写出矩阵 $\boldsymbol{M}_W\boldsymbol{X}$)

16.7 考虑回归模型：$Y_i=\beta_1X_i+\beta_2W_i+u_i$，这里为简单起见，省略了截距，且假定所有的变量都具有零均值。设 X_i 分布独立于(W_i,u_i)，但是 W_i 和 u_i 可能是相关的，并且设 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 为该模型的 OLS 估计量，证明：

a. 无论 W_i 和 u_i 是否相关，都有 $\hat{\beta}_1\xrightarrow{p}\beta_1$。

b. 如果 W_i 和 u_i 相关，$\hat{\beta}_2$ 是不一致的。

c. 设 $\hat{\beta}_1^r$ 是 Y 对 X 回归中(排除了 W 的那个有约束的回归)的 OLS 估计量，给出 $\hat{\beta}_1$ 具有比 $\hat{\beta}_1^r$ 更小的渐近方差的条件，允许 W_i 和 u_i 是相关的。

16.8 考虑回归模型 $Y_i=\beta_0+\beta_1X_i+u_i$，其中 $u_1=\tilde{u}_1$，且对于 $i=2,3,\cdots,n$，$u_i=0.5u_{i-1}+\tilde{u}_i$。设 $\tilde{u}_i$ 是独立同分布的，且均值为零，方差为 1，对于所有的 i 和 j，$\tilde{u}_i$ 与 X_j 是独立分布的。

*a. 推导 $E(\boldsymbol{UU}')=\boldsymbol{\Omega}$ 的表达式。

b. 请解释在没有明确地对矩阵 $\boldsymbol{\Omega}$ 进行转置的情况下，如何利用 GLS 估计该模型。(提示：变换该模型，使得回归误差为 $\tilde{u}_1,\tilde{u}_2,\cdots,\tilde{u}_n$)

16.9 这个练习说明：在条件均值独立性的假定(见附录 11.3)条件下，回归系数的一



个子集的 OLS 估计量是一致的。考虑矩阵形式的多元回归模型：$\boldsymbol{Y}=\boldsymbol{X\beta}+\boldsymbol{Wy}+\boldsymbol{u}$，其中，$\boldsymbol{X}$ 和 $\boldsymbol{W}$ 分别为回归因子的 $n\times k_1$ 和 $n\times k_2$ 阶矩阵。设 $\boldsymbol{X}_i'$ 和 $\boldsymbol{W}_i'$ 表示 $\boldsymbol{X}$ 和 $\boldsymbol{W}$ 的第 i 行（见公式(16.3)）。假设：(i) $E(u_i|\boldsymbol{X}_i,\boldsymbol{W}_i)=\boldsymbol{W}_i'\boldsymbol{\delta}$，其中，$\boldsymbol{\delta}$ 是未知参数的一个 $k_2\times 1$ 向量；(ii) $(\boldsymbol{X}_i,\boldsymbol{W}_i,Y_i)$ 是独立同分布的；(iii) $(\boldsymbol{X}_i,\boldsymbol{W}_i,u_i)$ 具有有限的非零的四阶矩；(iv) 不存在完全多重共线性。这就是重要概念 16.1 中的假设 1～4，只是用条件均值独立假设代替了通常的条件零均值假设。

a. 利用方程(16.45)表示 $\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}=(n^{-1}\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X})^{-1}(n^{-1}\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{U})$。

b. 证明：$n^{-1}\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{X}\xrightarrow{p}\boldsymbol{\Sigma}_{XX}-\boldsymbol{\Sigma}_{XW}\boldsymbol{\Sigma}_{WW}^{-1}\boldsymbol{\Sigma}_{WX}$，其中，$\boldsymbol{\Sigma}_{XX}=E(\boldsymbol{X}_i\boldsymbol{X}_i')$，$\boldsymbol{\Sigma}_{XW}=E(\boldsymbol{X}_i\boldsymbol{W}_i')$（如果对于所有的 i 和 j，$\boldsymbol{A}_{n,ij}\xrightarrow{p}\boldsymbol{A}_{ij}$，其中，$\boldsymbol{A}_{n,ij}$ 和 $\boldsymbol{A}_{ij}$ 分别是矩阵 $\boldsymbol{A}_n$ 和 $\boldsymbol{A}$ 的 (i,j) 元素，那么 $\boldsymbol{A}_n\xrightarrow{p}\boldsymbol{A}$）。

c. 证明：假设条件(i)和(ii)隐含着 $E(\boldsymbol{U}|\boldsymbol{X},\boldsymbol{W})=\boldsymbol{W\delta}$。

d. 利用(c)和累期望法则，证明：$n^{-1}\boldsymbol{X}'\boldsymbol{M}_W\boldsymbol{U}\xrightarrow{p}\boldsymbol{0}_{k_1\times 1}$。

e. 利用(a)～(d)证明：在满足(i)～(iv)的条件下，$\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$。

附录 16.1 矩阵代数简介

本附录概述了第 16 章中所用到的向量、矩阵以及矩阵代数的其他一些基本内容。本附录的目的是复习线性代数这门课程中的一些基本概念和定义，而不是完全复述这门课程。

向量与矩阵的定义

向量（vector）是由 n 个数或元素组成的一个汇集，以列的方式汇集称为列向量（column vector），以行的方式汇集称为**行向量**（row vector）。n 维列向量 $\boldsymbol{b}$ 和 n 维行向量 $\boldsymbol{c}$ 记为：

$$\boldsymbol{b}=\begin{bmatrix}b_1\\b_2\\\vdots\\b_n\end{bmatrix},\boldsymbol{c}=[c_1\ c_2\cdots c_n]$$

其中，b_1 是 $\boldsymbol{b}$ 的第一个元素，b_i 是 $\boldsymbol{b}$ 的第 i 个元素。

整本书中，我们都是用黑体符号表示向量和矩阵。

矩阵（matrix）是数字或元素的汇集或排列，各元素以行和列进行排列。一个矩阵的阶数是 $n\times m$，其中 n 表示行数，m 表示列数。$n\times m$ 阶矩阵 $\boldsymbol{A}$ 为：

$$\boldsymbol{A}=\begin{bmatrix}a_{11}&a_{12}&\cdots&a_{1m}\\a_{21}&a_{22}&\cdots&a_{2m}\\\vdots&\vdots&&\vdots\\a_{n1}&a_{n2}&\cdots&a_{nm}\end{bmatrix}$$

其中，a_{ij} 是 $\boldsymbol{A}$ 的 (i,j) 元素，即 a_{ij} 是出现在第 i 行、第 j 列的元素。$n\times m$ 阶矩阵由 n 个行向量或 m 个列向量构成。

为了将一维数值与向量或矩阵区别开来，我们称一维数值为**标量**（scalar）。

g. $(\boldsymbol{AB})' = \boldsymbol{B}'\boldsymbol{A}'$。

一般地说，矩阵相乘不能交换次序，即 $\boldsymbol{AB} \neq \boldsymbol{BA}$，只有在一些特殊情况下矩阵相乘才可以交换次序，例如，当 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都是 $n \times n$ 阶对角阵时，有 $\boldsymbol{AB} = \boldsymbol{BA}$。

矩阵逆、矩阵方根及其相关问题

矩阵的逆。设 $\boldsymbol{A}$ 为方阵。假设矩阵 $\boldsymbol{A}$ 的逆（inverse）$\boldsymbol{A}^{-1}$ 存在，它被定义为满足 $\boldsymbol{A}^{-1}\boldsymbol{A} = \boldsymbol{I}$ 的矩阵。实际上，如果逆矩阵 $\boldsymbol{A}^{-1}$ 存在，那么 $\boldsymbol{A}$ 被称为**可逆的**（invertible）或**非奇异的**（nonsigular）。如果 $\boldsymbol{A}$ 和 $\boldsymbol{B}$ 都可逆，那么 $(\boldsymbol{AB})^{-1} = \boldsymbol{B}^{-1}\boldsymbol{A}^{-1}$。

正定和半正定阵。设 $\boldsymbol{V}$ 为 $n \times n$ 方阵。如果对于任意非零的 $n \times 1$ 向量 $\boldsymbol{c}$，有 $\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c} > 0$，则称 $\boldsymbol{V}$ 是**正定的**（positive definite）。同样，如果对于任意非零的 $n \times 1$ 向量 $\boldsymbol{c}$，有 $\boldsymbol{c}'\boldsymbol{V}\boldsymbol{c} \geqslant 0$，则称 $\boldsymbol{V}$ 是**半正定的**（positive semidefinite）。如果 $\boldsymbol{V}$ 是正定的，那么它是可逆的。

线性独立。如果不存在非零的标量 c_1 和 c_2，使得 $c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 = 0_{n \times 1}$，那么 $n \times 1$ 向量 $\boldsymbol{a}_1$ 和 $\boldsymbol{a}_2$ 是**线性独立的**（linearly independent）。更一般地说，如果不存在非零标量 $c_1, c_2, \cdots, c_k$ 使得 $c_1\boldsymbol{a}_1 + c_2\boldsymbol{a}_2 + \cdots + c_k\boldsymbol{a}_k = \boldsymbol{0}_{n \times 1}$，那么称 k 个向量 $\boldsymbol{a}_1, \boldsymbol{a}_2, \cdots, \boldsymbol{a}_k$ 的集合为线性独立的。

矩阵的秩。$n \times m$ 阶矩阵 $\boldsymbol{A}$ 的秩（rank）是线性独立的列向量的个数，$\boldsymbol{A}$ 的秩记为 $rank(\boldsymbol{A})$。如果 $\boldsymbol{A}$ 的秩等于 $\boldsymbol{A}$ 的列数，则称其为列满秩的。如果 $n \times m$ 矩阵 $\boldsymbol{A}$ 是列满秩的，则不存在一个非零的 $m \times 1$ 向量 $\boldsymbol{c}$，使得 $\boldsymbol{Ac} = \boldsymbol{0}_{n \times 1}$。如果 $\boldsymbol{A}$ 为 $n \times n$ 方阵，且 $rank(\boldsymbol{A}) = n$，那么 $\boldsymbol{A}$ 是非奇异的。如果 $n \times m$ 矩阵 $\boldsymbol{A}$ 是列满秩的，则 $\boldsymbol{A}'\boldsymbol{A}$ 是非奇异的。

矩阵平方根。设 $\boldsymbol{V}$ 为 $n \times n$ 阶对称正定方阵。$\boldsymbol{V}$ 的平方根阵定义为 $n \times n$ 阶矩阵 $\boldsymbol{F}$，使得满足 $\boldsymbol{F}'\boldsymbol{F} = \boldsymbol{V}$。一个正定矩阵的平方根总是存在的，但是它不是惟一的。矩阵平方根具有 $\boldsymbol{F}\boldsymbol{V}^{-1}\boldsymbol{F}' = \boldsymbol{I}_n$ 的性质。此外，一个正定矩阵的平方根是可逆的，所以，$\boldsymbol{F}'^{-1}\boldsymbol{V}\boldsymbol{F}^{-1} = \boldsymbol{I}_n$。

附录 16.2　多变量分布

本附录汇集了关于随机变量向量分布的几个定义及事实。我们从定义 n 维随机变量 $\boldsymbol{V}$ 的均值和协方差矩阵开始，接下来给出了多元正态分布，最后归纳了关于联合正态分布随机变量的线性函数和二次函数的一些事实。

均值向量和方差矩阵

随机变量的一个 $m \times 1$ 向量 $\boldsymbol{V} = (V_1 V_2 \cdots V_m)'$ 的一阶矩和二阶矩，由它们的均值向量和协方差阵来概括。

由于 $\boldsymbol{V}$ 是一个向量，因此它的均值也是个向量，即它的**均值向量**（mean vector）是 $E(\boldsymbol{V}) = \boldsymbol{\mu}_V$。均值向量的第 i 个元素是 $\boldsymbol{V}$ 的第 i 个元素的均值。

$\boldsymbol{V}$ 的**协方差矩阵**（covariance matrix），是由沿着对角线的方差 $\text{var}(V_i)$，$i = 1, \cdots, n$ 和 (i,j) 非对角线上的元素 $\text{cov}(V_i, V_j)$ 组成的矩阵。写成矩阵形式，协方差矩阵 $\boldsymbol{\Sigma}_V$ 为：

$$\boldsymbol{\Sigma}_V = E[(\boldsymbol{V} - \boldsymbol{\mu}_V)(\boldsymbol{V} - \boldsymbol{\mu}_V)'] = \begin{pmatrix} \text{var}(V_1) & \cdots & \text{cov}(V_1, V_m) \\ \vdots & \ddots & \vdots \\ \text{cov}(V_m, V_1) & \cdots & \text{var}(V_m) \end{pmatrix} \tag{16.46}$$



多元正态分布

如果 $m \times 1$ 向量随机变量 $\boldsymbol{V}$ 具有联合概率密度函数：

$$f(V)=\frac{1}{\sqrt{(2\pi)^m\det(\boldsymbol{\Sigma}_V)}}\exp[-\frac{1}{2}(V-\boldsymbol{\mu}_V)'\boldsymbol{\Sigma}_V^{-1}(V-\boldsymbol{\mu}_V)] \tag{16.47}$$

那么它服从均值为向量$\boldsymbol{\mu}_V$、协方差矩阵为$\boldsymbol{\Sigma}_V$的多元正态分布，其中，$\det(\boldsymbol{\Sigma}_V)$是矩阵$\boldsymbol{\Sigma}_V$的行列式。多元正态分布表示为$N(\boldsymbol{\mu}_V,\boldsymbol{\Sigma}_V)$。

关于多元正态分布的一个重要的事实是：如果两个联合正态分布的随机变量不相关（或者说，具有零协方差矩阵），那么它们是独立分布的。也就是说，设V_1和V_2为联合正态分布随机变量，维数分别为$m_1\times1$和$m_2\times1$，如果$\text{cov}(V_1,V_2)=E[(V_1-\boldsymbol{\mu}_{V_1})(V_2-\boldsymbol{\mu}_{V_2})']=\mathbf{0}_{m_1\times m_2}$，那么$V_1$和$V_2$是独立的。

如果$\{V_i\}$是独立同分布的$N(0,\sigma_v^2)$，那么$\boldsymbol{\Sigma}_V=\sigma_v^2\boldsymbol{I}_m$，而且多元正态分布简化为$m$个单变量的正态密度的乘积。

正态随机变量的线性组合和二次型的分布

多元正态随机变量的线性组合仍然是正态分布的，多元正态随机变量的某些二次型具有卡方分布。设V是一个$m\times1$的随机变量，服从分布$N(\boldsymbol{\mu}_V,\boldsymbol{\Sigma}_V)$，设$\boldsymbol{A}$和$\boldsymbol{B}$分别为非随机的$a\times m$和$b\times m$阶矩阵，设$\boldsymbol{d}$为非随机的$a\times1$向量，那么：

$$\boldsymbol{d}+\boldsymbol{A}V\sim N(\boldsymbol{d}+\boldsymbol{A}\boldsymbol{\mu}_V,\boldsymbol{A}\boldsymbol{\Sigma}_V\boldsymbol{A}') \tag{16.48}$$

$$\text{cov}(\boldsymbol{A}V,\boldsymbol{B}V)=\boldsymbol{A}\boldsymbol{\Sigma}_V\boldsymbol{B}' \tag{16.49}$$

$$\text{如果}\boldsymbol{A}\boldsymbol{\Sigma}_V\boldsymbol{B}'=\mathbf{0}_{a\times b}\text{，那么}\boldsymbol{A}V\text{和}\boldsymbol{B}V\text{是独立分布的} \tag{16.50}$$

$$V'\boldsymbol{\Sigma}_V^{-1}V\sim\chi_m^2 \tag{16.51}$$

关于二次型的另一个有用的结果涉及幂等阵。如果矩阵$\boldsymbol{C}$是方阵，且$\boldsymbol{CC}=\boldsymbol{C}$，则矩阵$\boldsymbol{C}$是幂等的。设$V$是一个$m$维多元正态随机变量，其分布为$N(\mathbf{0},\sigma_v^2\boldsymbol{I}_m)$，如果$\boldsymbol{C}$是对称的且是幂等的，那么：

$$V'\boldsymbol{C}V/\sigma_v^2\sim\chi_r^2\text{，其中，}r=rank(\boldsymbol{C}) \tag{16.52}$$

附录16.3　$\hat{\boldsymbol{\beta}}$的渐近分布的推导

本附录给出了表达式(16.12)中$\sqrt{n}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$的渐近正态分布的推导。这个结果的一个含义是$\hat{\boldsymbol{\beta}}\xrightarrow{p}\boldsymbol{\beta}$。

首先，考虑公式(16.15)中的分母矩阵$\boldsymbol{X}'\boldsymbol{X}/n=\frac{1}{n}\sum_{i=1}^{n}\boldsymbol{X}_i\boldsymbol{X}_i'$。这个矩阵的$(j,l)$元素为$\frac{1}{n}\sum_{i=1}^{n}X_{ji}X_{li}$。根据重要概念16.1中的第二个假设，$\boldsymbol{X}_i$是独立同分布的，从而$X_{ji}X_{li}$是独立同分布的。根据重要概念16.1中的第三个假设，$\boldsymbol{X}_i$的每一个元素都具有四阶矩，所以，根据Cauchy-Schwarz不等式（见附录15.2），$X_{ji}X_{li}$具有二阶矩。由于$X_{ji}X_{li}$有二阶矩且独立同分布，$\frac{1}{n}\sum_{i=1}^{n}X_{ji}X_{li}$遵从大数定律，因此$\frac{1}{n}\sum_{i=1}^{n}X_{ji}X_{li}\xrightarrow{p}E(X_{ji}X_{li})$。这对$\boldsymbol{X}'\boldsymbol{X}/n$的所有元素都成立，所以$\boldsymbol{X}'\boldsymbol{X}/n\xrightarrow{p}E(\boldsymbol{X}_i\boldsymbol{X}_i')=\boldsymbol{Q}_X$。



其次，考虑公式(16.15)中的分子矩阵，$\boldsymbol{X}'\boldsymbol{U}/\sqrt{n}=\sqrt{\frac{1}{n}}\sum_{i=1}^{n}\boldsymbol{V}_i$，其中$\boldsymbol{V}_i=\boldsymbol{X}_iu_i$。根据重要概念16.1中的第一个假设和累期望法则，$E(\boldsymbol{V}_i)=E[\boldsymbol{X}_iE(u_i|\boldsymbol{X}_i)]=\mathbf{0}_{k+1}$。根据第二个最

小二乘假设，V_t 是独立同分布的。设 c 是一个有限的 $k+1$ 维向量，根据 Cauchy-Schwarz 不等式，$E[(c'V_t)^2]=E[(c'X_tu_t)^2]=E[(c'X_t)^2(u_t)^2]\leqslant\sqrt{E[(c'X_t)^4]E(u_t^4)}$，根据第三个最小二乘假设，它是有限的。对于任意的向量 c，这个结果都成立，所以 $E(V_tV_t')=\Sigma_V$ 也是有限的，并且我们假定它是正定的。因此，重要概念 16.2 中的多元中心极限定理适用于 $\sqrt{\frac{1}{n}}\sum_{t=1}^{n}V_t=\frac{1}{\sqrt{n}}X'U$，即：

$$\frac{1}{\sqrt{n}}X'U\xrightarrow{d}N(0_{k+1},\Sigma_V) \tag{16.53}$$

公式(16.12)中的结果是根据公式(16.15)、表达式(16.53)、$X'X/n$ 的一致性、第四个最小二乘假设(确保了 $(X'X)^{-1}$ 的存在)以及 Slutsky 定理得到的。

附录 16.4 含有正态误差的 OLS 检验统计量的精确分布的推导

本附录给出了表达式(16.35)中的仅适用于同方差的 t 统计量和表达式(16.37)中的仅适用于同方差的 F 统计量，在零假设下它们的分布的证明，假设重要概念 16.1 中的所有六个假设都成立。

表达式(16.35)的证明

如果：(i) Z 服从标准正态分布，(ii) W 服从 χ_m^2 分布，(iii) Z 和 W 是独立分布的，那么随机变量 $Z/\sqrt{W/m}$ 服从自由度为 m 的 t 分布(见附录 15.1)。注意 $\widetilde{\Sigma}_{\hat{\beta}}=(s_{\hat{u}}^2/\sigma_u^2)\Sigma_{\hat{\beta}|X}$，那么表达式(16.34)可重写成：

$$\tilde{t}=\frac{(\hat{\beta}_j-\beta_{j,0})/\sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}}{\sqrt{W/(n-k-1)}} \tag{16.54}$$

其中，$W=(n-k-1)(s_{\hat{u}}^2/\sigma_u^2)$，并且令 $Z=(\hat{\beta}_j-\beta_{j,0})/\sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}$，$m=n-k-1$。在这些定义下，$\tilde{t}=Z/\sqrt{W/m}$。因此，要证明表达式(16.35)中的结果，我们必须证明对于 Z，W 和 m 的这些定义，(i)～(iii)的结论成立。

i. 表达式(16.30)的含义是：在零假设下，$Z=(\hat{\beta}_j-\beta_{j,0})/\sqrt{(\Sigma_{\hat{\beta}|X})_{jj}}$ 服从精确的标准正态分布，这就证明了(i)。

ii. 根据表达式(16.31)，W 服从 χ_{n-k-1}^2 分布，这就证明了(ii)。

iii. 要证明(iii)成立，必须证明 $\hat{\beta}_j$ 和 $s_{\hat{u}}^2$ 是独立分布的。

根据公式(16.14)和公式(16.29)，$\hat{\beta}-\beta=(X'X)^{-1}X'U$，且 $s_{\hat{u}}^2=(M_XU)'(M_XU)/(n-k-1)$。因此，如果 $(X'X)^{-1}X'U$ 和 M_XU 独立，那么 $\hat{\beta}-\beta$ 和 $s_{\hat{u}}^2$ 独立。$(X'X)^{-1}X'U$ 和 M_XU 都是 U 的线性组合，它服从分布 $N(0_{n\times1},\sigma_u^2I_n)$，以 X 为条件。但是，因为 $M_XX(X'X)^{-1}=0_{n\times(k+1)}$(见表达式(16.26))，由此得出 $(X'X)^{-1}X'U$ 和 M_XU 是独立分布的(见表达式(16.50))。因此，在重要概念 16.1 的所有六个假设下：

$$\hat{\beta}\text{ 和 }s_{\hat{u}}^2\text{ 是独立分布的} \tag{16.55}$$

这就证明了(iii)，从而也证明了表达式(16.35)。

第5部分



表达式(16.37)的证明

F_{n_1,n_2}的分布就是$(W_1/n_1)/(W_2/n_2)$的分布，其中，(i)W_1服从$\chi^2_{n_1}$分布；(ii)W_2服从$\chi^2_{n_2}$分布；(iii)W_1和W_2是独立分布的(见附录15.1)。为了用$\widetilde{F}$形式表达，设$W_1=(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma_u^2]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$，$W_2=(n-k-1)s_{\hat{u}}^2/\sigma_u^2$。将这些定义代入公式(16.36)，得到$\widetilde{F}=(W_1/q)/[W_2/(n-k-1)]$。因此，根据$F$分布的定义，如果(i)~(iii)成立，且$n_1=q$，$n_2=n-k-1$，则$\widetilde{F}$服从$F_{q,n-k-1}$分布。

i. 在零假设下，$\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}=\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$。因为在表达式(16.30)中，$\hat{\boldsymbol{\beta}}$服从条件正态分布，$\boldsymbol{R}$是一个非随机矩阵，所以，以$\boldsymbol{X}$为条件的$\boldsymbol{R}(\hat{\boldsymbol{\beta}}-\boldsymbol{\beta})$服从分布$N(\boldsymbol{0}_{q\times1},\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma_u^2)$。因此，根据附录16.2中的表达式(16.51)，$(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})'[\boldsymbol{R}(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{R}'\sigma_u^2]^{-1}(\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r})$服从$\chi_q^2$分布，从而证明了(i)。

ii. (ii)中的要求已由表达式(16.31)给出。

iii. 已证明$\hat{\boldsymbol{\beta}}-\boldsymbol{\beta}$和$s_{\hat{u}}^2$是独立分布的(见表达式(16.55))，从而$\boldsymbol{R}\hat{\boldsymbol{\beta}}-\boldsymbol{r}$和$s_{\hat{u}}^2$是独立分布的，这反过来又意味着$W_1$和$W_2$是独立分布的，这就证明了(iii)。证毕。

附录16.5　多元回归的高斯—马尔可夫定理的证明

本附录证明了多元回归模型的高斯—马尔可夫定理(见重要概念16.3)。设$\widetilde{\boldsymbol{\beta}}$是$\boldsymbol{\beta}$的一个线性条件无偏估计量，于是$\widetilde{\boldsymbol{\beta}}=\boldsymbol{A}'\boldsymbol{Y}$且$E(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})=\boldsymbol{\beta}$，其中，$\boldsymbol{A}$是一个$n\times(k+1)$阶矩阵，它依赖于$\boldsymbol{X}$和非随机常数。我们证明，对于所有的$k+1$维向量$\boldsymbol{c}$，有$\mathrm{var}(\boldsymbol{c}'\hat{\boldsymbol{\beta}})\leqslant\mathrm{var}(\boldsymbol{c}'\widetilde{\boldsymbol{\beta}})$。这里当且仅当$\widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}$时，不等式中的等式才成立。

由于$\widetilde{\boldsymbol{\beta}}$是线性的，因此，它可被写为$\widetilde{\boldsymbol{\beta}}=\boldsymbol{A}'\boldsymbol{Y}=\boldsymbol{A}'(\boldsymbol{X\beta}+\boldsymbol{U})=(\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta}+\boldsymbol{A}'\boldsymbol{U}$。根据第一个高斯—马尔可夫条件，$E(\boldsymbol{U}|\boldsymbol{X})=\boldsymbol{0}_{n\times1}$，所以，$E(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})=(\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta}$，但是，由于$\widetilde{\boldsymbol{\beta}}$是条件无偏的$E(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})=\boldsymbol{\beta}=(\boldsymbol{A}'\boldsymbol{X})\boldsymbol{\beta}$，这意味着$\boldsymbol{A}'\boldsymbol{X}=\boldsymbol{I}_{k+1}$，因此$\widetilde{\boldsymbol{\beta}}=\boldsymbol{\beta}+\boldsymbol{A}'\boldsymbol{U}$，从而$\mathrm{var}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})=\mathrm{var}(\boldsymbol{A}'\boldsymbol{U}|\boldsymbol{X})=E(\boldsymbol{A}'\boldsymbol{U}\boldsymbol{U}'\boldsymbol{A}|\boldsymbol{X})=\boldsymbol{A}'E(\boldsymbol{U}\boldsymbol{U}'|\boldsymbol{X})\boldsymbol{A}=\sigma_u^2\boldsymbol{A}'\boldsymbol{A}$，其中，第三个等式成立是因为$\boldsymbol{A}$会依赖于$\boldsymbol{X}$但不依赖于$\boldsymbol{U}$，最后一个等式成立是因为第二个高斯—马尔可夫条件，即如果$\widetilde{\boldsymbol{\beta}}$是线性的和无偏的，那么在高斯—马尔可夫条件下：

$$\boldsymbol{A}'\boldsymbol{X}=\boldsymbol{I}_{k+1},\mathrm{var}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})=\sigma_u^2\boldsymbol{A}'\boldsymbol{A} \tag{16.56}$$

当$\boldsymbol{A}=\hat{\boldsymbol{A}}=\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}$时，等式(16.56)中的结果也适用于$\hat{\boldsymbol{\beta}}$，这里，根据第三个高斯—马尔可夫条件，$(\boldsymbol{X}'\boldsymbol{X})^{-1}$存在。

现在设$\boldsymbol{A}=\hat{\boldsymbol{A}}+\boldsymbol{D}$，因此，$\boldsymbol{D}$为权数矩阵$\boldsymbol{A}$和$\hat{\boldsymbol{A}}$之差。注意，因为$\hat{\boldsymbol{A}}'\boldsymbol{A}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{A}=(\boldsymbol{X}'\boldsymbol{X})^{-1}$(根据公式(16.56))，$\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}\boldsymbol{X}'\boldsymbol{X}(\boldsymbol{X}'\boldsymbol{X})^{-1}=(\boldsymbol{X}'\boldsymbol{X})^{-1}$，所以，$\hat{\boldsymbol{A}}'\boldsymbol{D}=\hat{\boldsymbol{A}}'(\boldsymbol{A}-\hat{\boldsymbol{A}})=\hat{\boldsymbol{A}}'\boldsymbol{A}-\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}}=\boldsymbol{0}_{(k+1)\times(k+1)}$。将$\boldsymbol{A}=\hat{\boldsymbol{A}}+\boldsymbol{D}$代入公式(16.56)中的条件方差公式，得到：

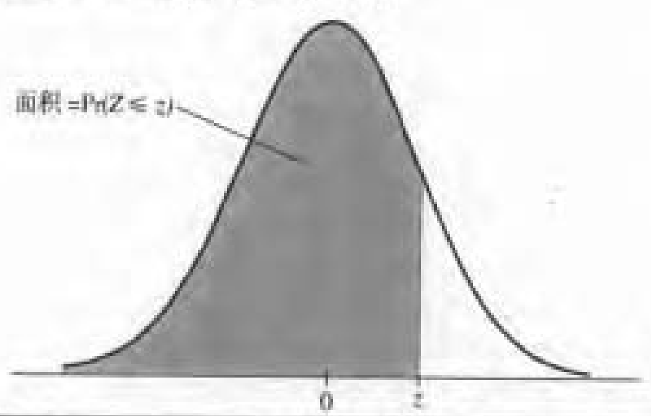
$$\begin{aligned}\mathrm{var}(\widetilde{\boldsymbol{\beta}}|\boldsymbol{X})&=\sigma_u^2(\hat{\boldsymbol{A}}+\boldsymbol{D})'(\hat{\boldsymbol{A}}+\boldsymbol{D})\\&=\sigma_u^2[\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}}+\hat{\boldsymbol{A}}'\boldsymbol{D}+\boldsymbol{D}'\hat{\boldsymbol{A}}+\boldsymbol{D}'\boldsymbol{D}]\\&=\sigma_u^2(\boldsymbol{X}'\boldsymbol{X})^{-1}+\sigma_u^2\boldsymbol{D}'\boldsymbol{D}\end{aligned} \tag{16.57}$$

其中，最后一个等式利用了$\hat{\boldsymbol{A}}'\hat{\boldsymbol{A}}=(\boldsymbol{X}'\boldsymbol{X})^{-1}$和$\hat{\boldsymbol{A}}'\boldsymbol{D}'=\boldsymbol{0}_{(k+1)\times(k+1)}$的事实。



附录

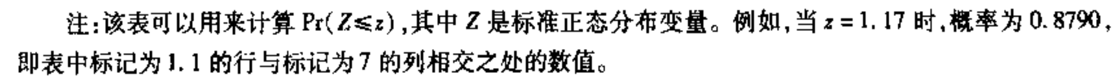
附表1　累积标准正态分布函数，$\phi(z)=\Pr(Z\leq z)$



z	z的第二个小数位的值									
	0	1	2	3	4	5	6	7	8	9
-2.9	0.0019	0.0018	0.0018	0.0017	0.0016	0.0016	0.0015	0.0015	0.0014	0.0014
-2.8	0.0026	0.0025	0.0024	0.0023	0.0023	0.0022	0.0021	0.0021	0.0020	0.0019
-2.7	0.0035	0.0034	0.0033	0.0032	0.0031	0.0030	0.0029	0.0028	0.0027	0.0026
-2.6	0.0047	0.0045	0.0044	0.0043	0.0041	0.0040	0.0039	0.0038	0.0037	0.0036
-2.5	0.0062	0.0060	0.0059	0.0057	0.0055	0.0054	0.0052	0.0051	0.0049	0.0048
-2.4	0.0082	0.0080	0.0078	0.0075	0.0073	0.0071	0.0069	0.0068	0.0066	0.0064
-2.3	0.0107	0.0104	0.0102	0.0099	0.0096	0.0094	0.0091	0.0089	0.0087	0.0084
-2.2	0.0139	0.0136	0.0132	0.0129	0.0125	0.0122	0.0119	0.0116	0.0113	0.0110
-2.1	0.0179	0.0174	0.0170	0.0166	0.0162	0.0158	0.0154	0.0150	0.0146	0.0143
-2.0	0.0228	0.0222	0.0217	0.0212	0.0207	0.0202	0.0197	0.0192	0.0188	0.0183
-1.9	0.0287	0.0281	0.0274	0.0268	0.0262	0.0256	0.0250	0.0244	0.0239	0.0233
-1.8	0.0359	0.0351	0.0344	0.0336	0.0329	0.0322	0.0314	0.0307	0.0301	0.0294
-1.7	0.0446	0.0436	0.0427	0.0418	0.0409	0.0401	0.0392	0.0384	0.0375	0.0367
-1.6	0.0548	0.0537	0.0526	0.0516	0.0505	0.0495	0.0485	0.0475	0.0465	0.0455
-1.5	0.0668	0.0655	0.0643	0.0630	0.0618	0.0606	0.0594	0.0582	0.0571	0.0559
-1.4	0.0808	0.0793	0.0778	0.0764	0.0749	0.0735	0.0721	0.0708	0.0694	0.0681
-1.3	0.0968	0.0951	0.0934	0.0918	0.0901	0.0885	0.0869	0.0853	0.0838	0.0823
-1.2	0.1151	0.1131	0.1112	0.1093	0.1075	0.1056	0.1038	0.1020	0.1003	0.0985
-1.1	0.1357	0.1335	0.1314	0.1292	0.1271	0.1251	0.1230	0.1210	0.1190	0.1170
-1.0	0.1587	0.1562	0.1539	0.1515	0.1492	0.1469	0.1446	0.1423	0.1401	0.1379
-0.9	0.1841	0.1814	0.1788	0.1762	0.1736	0.1711	0.1685	0.1660	0.1635	0.1611

续表

z	Z的第二个小数位的值									
	0	1	2	3	4	5	6	7	8	9
-0.8	0.2119	0.2090	0.2061	0.2033	0.2005	0.1977	0.1949	0.1922	0.1894	0.1867
-0.7	0.2420	0.2389	0.2358	0.2327	0.2296	0.2266	0.2236	0.2206	0.2177	0.2148
-0.6	0.2743	0.2709	0.2676	0.2643	0.2611	0.2578	0.2546	0.2514	0.2483	0.2451
-0.5	0.3085	0.3050	0.3015	0.2981	0.2946	0.2912	0.2877	0.2843	0.2810	0.2776
-0.4	0.3446	0.3409	0.3372	0.3336	0.3300	0.3264	0.3228	0.3192	0.3156	0.3121
-0.3	0.3821	0.3783	0.3745	0.3707	0.3669	0.3632	0.3594	0.3557	0.3520	0.3483
-0.2	0.4207	0.4168	0.4129	0.4090	0.4052	0.4013	0.3974	0.3936	0.3897	0.3859
-0.1	0.4602	0.4562	0.4522	0.4483	0.4443	0.4404	0.4364	0.4325	0.4286	0.4247
-0.0	0.5000	0.4960	0.4920	0.4880	0.4840	0.4801	0.4761	0.4721	0.4681	0.4641
0.0	0.5000	0.5040	0.5080	0.5120	0.5160	0.5199	0.5239	0.5279	0.5319	0.5359
0.1	0.5398	0.5438	0.5478	0.5517	0.5557	0.5596	0.5636	0.5675	0.5714	0.5753
0.2	0.5793	0.5832	0.5871	0.5910	0.5948	0.5987	0.6026	0.6064	0.6103	0.6141
0.3	0.6179	0.6217	0.6255	0.6293	0.6331	0.6368	0.6406	0.6443	0.6480	0.6517
0.4	0.6554	0.6591	0.6628	0.6664	0.6700	0.6736	0.6772	0.6808	0.6844	0.6879
0.5	0.6915	0.6950	0.6985	0.7019	0.7054	0.7088	0.7123	0.7157	0.7190	0.7224
0.6	0.7257	0.7291	0.7324	0.7357	0.7389	0.7422	0.7454	0.7486	0.7517	0.7549
0.7	0.7580	0.7611	0.7642	0.7673	0.7704	0.7734	0.7764	0.7794	0.7823	0.7852
0.8	0.7881	0.7910	0.7939	0.7967	0.7995	0.8023	0.8051	0.8078	0.8106	0.8133
0.9	0.8159	0.8186	0.8212	0.8238	0.8264	0.8289	0.8315	0.8340	0.8365	0.8389
1.0	0.8413	0.8438	0.8461	0.8485	0.8508	0.8531	0.8554	0.8577	0.8599	0.8621
1.1	0.8643	0.8665	0.8686	0.8708	0.8729	0.8749	0.8770	0.8790	0.8810	0.8830
1.2	0.8849	0.8869	0.8888	0.8907	0.8925	0.8944	0.8962	0.8980	0.8997	0.9015
1.3	0.9032	0.9049	0.9066	0.9082	0.9099	0.9115	0.9131	0.9147	0.9162	0.9711
1.4	0.9192	0.9207	0.9222	0.9236	0.9251	0.9265	0.9279	0.9292	0.9306	0.9319
1.5	0.9332	0.9345	0.9357	0.9370	0.9382	0.9394	0.9406	0.9418	0.9429	0.9441
1.6	0.9452	0.9463	0.9474	0.9484	0.9495	0.9505	0.9515	0.9525	0.9535	0.9545
1.7	0.9554	0.9564	0.9573	0.9582	0.9591	0.9599	0.9608	0.9616	0.9625	0.9633
1.8	0.9641	0.9649	0.9656	0.9664	0.9671	0.9678	0.9686	0.9693	0.9699	0.9706
1.9	0.9713	0.9719	0.9726	0.9732	0.9738	0.9744	0.9750	0.9756	0.9761	0.9767
2.0	0.9772	0.9778	0.9783	0.9788	0.9793	0.9798	0.9803	0.9808	0.9812	0.9817
2.1	0.9821	0.9826	0.9830	0.9834	0.9838	0.9842	0.9846	0.9850	0.9854	0.9857
2.2	0.9861	0.9864	0.9868	0.9871	0.9875	0.9878	0.9881	0.9884	0.9887	0.9890
2.3	0.9893	0.9896	0.9898	0.9901	0.9904	0.9906	0.9909	0.9911	0.9913	0.9916
2.4	0.9918	0.9920	0.9922	0.9925	0.9927	0.9929	0.9931	0.9932	0.9934	0.9936
2.5	0.9938	0.9940	0.9941	0.9943	0.9945	0.9946	0.9948	0.9949	0.9951	0.9952
2.6	0.9953	0.9955	0.9956	0.9957	0.9959	0.9960	0.9961	0.9962	0.9963	0.9964
2.7	0.9965	0.9966	0.9967	0.9968	0.9969	0.9970	0.9971	0.9972	0.9973	0.9974
2.8	0.9974	0.9975	0.9976	0.9977	0.9977	0.9978	0.9979	0.9979	0.9980	0.9981
2.9	0.9981	0.9982	0.9982	0.9983	0.9984	0.9984	0.9985	0.9985	0.9986	0.9986

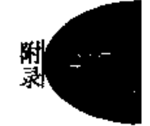

注：该表可以用来计算 $\Pr(Z \leqslant z)$，其中 Z 是标准正态分布变量。例如，当 $z=1.17$ 时，概率为 0.8790，即表中标记为 1.1 的行与标记为 7 的列相交之处的数值。



附表 2　双边和单边检验的学生 t 分布临界值

自由度	显著性水平				
	20%（双边） 10%（单边）	10%（双边） 5%（单边）	5%（双边） 2.5%（单边）	2%（双边） 1%（单边）	1%（双边） 0.5%（单边）
1	3.08	6.31	12.71	31.82	63.66
2	1.89	2.92	4.30	6.96	9.92
3	1.64	2.35	3.18	4.54	5.84
4	1.53	2.13	2.78	3.75	4.60
5	1.48	2.02	2.57	3.36	4.03
6	1.44	1.94	2.45	3.14	3.71
7	1.41	1.89	2.36	3.00	3.50
8	1.40	1.86	2.31	2.90	3.36
9	1.38	1.83	2.26	2.82	3.25
10	1.37	1.81	2.23	2.76	3.17
11	1.36	1.80	2.20	2.72	3.11
12	1.36	1.78	2.18	2.68	3.05
13	1.35	1.77	2.16	2.65	3.01
14	1.35	1.76	2.14	2.62	2.98
15	1.34	1.75	2.13	2.60	2.95
16	1.34	1.75	2.12	2.58	2.92
17	1.33	1.74	2.11	2.57	2.90
18	1.33	1.73	2.10	2.55	2.88
19	1.33	1.73	2.09	2.54	2.86
20	1.33	1.72	2.09	2.53	2.85
21	1.32	1.72	2.08	2.52	2.83
22	1.32	1.72	2.07	2.51	2.82
23	1.32	1.71	2.07	2.50	2.81
24	1.32	1.71	2.06	2.49	2.80
25	1.32	1.71	2.06	2.49	2.79
26	1.32	1.71	2.06	2.48	2.78
27	1.31	1.70	2.05	2.47	2.77
28	1.31	1.70	2.05	2.47	2.76
29	1.31	1.70	2.05	2.46	2.76
30	1.31	1.70	2.04	2.46	2.75
60	1.30	1.67	2.00	2.39	2.66
90	1.29	1.66	1.99	2.37	2.63
120	1.29	1.66	1.98	2.36	2.62
∞	1.28	1.64	1.96	2.33	2.58

注：表中所示为备择假设的双边（≠）和单边（>）的临界值。单边检验（<）检验的临界值为表中所示的单边检验（>）的临界值的相反数。例如，2.13 为利用自由度为 15 的学生 t 分布的双边检验的 5% 的显著性水平的临界值。

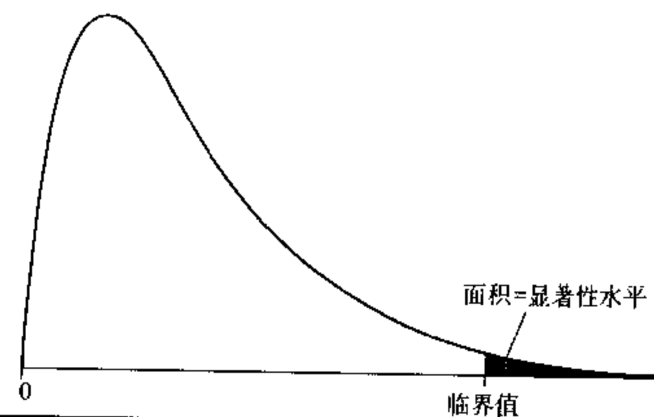


附表 3　卡方分布的临界值

自由度	显著性水平		
	10%	5%	1%
1	2.71	3.84	6.63
2	4.61	5.99	9.21
3	6.25	7.81	11.34
4	7.78	9.49	13.28
5	9.24	11.07	15.09
6	10.64	12.59	16.81
7	12.02	14.07	18.48
8	13.36	15.51	20.09
9	14.68	16.92	21.67
10	15.99	18.31	23.21
11	17.28	19.68	24.72
12	18.55	21.03	26.22
13	19.81	22.36	27.69
14	21.06	23.68	29.14
15	22.31	25.00	30.58
16	23.54	26.30	32.00
17	24.77	27.59	33.41
18	25.99	28.87	34.81
19	27.20	30.14	36.19
20	28.41	31.41	37.57
21	29.62	32.67	38.93
22	30.81	33.92	40.29
23	32.01	35.17	41.64
24	33.20	36.41	42.98
25	34.38	37.65	44.31
26	35.56	38.89	45.64
27	36.74	40.11	46.96
28	37.92	41.34	48.28
29	39.09	42.56	49.59
30	40.26	43.77	50.89

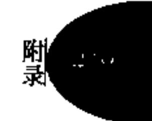
注:该表给出了卡方分布的第 90、第 95 和第 99 个百分位数。它们可作为在 10%,5% 和 1% 的显著性水平下检验的临界值。

附表 4　$F_{m,\infty}$ 分布的临界值



自由度	显著性水平		
	10%	5%	1%
1	2.71	3.84	6.63
2	2.30	3.00	4.61
3	2.08	2.60	3.78
4	1.94	2.37	3.32
5	1.85	2.21	3.02
6	1.77	2.10	2.80
7	1.72	2.01	2.64
8	1.67	1.94	2.51
9	1.63	1.88	2.41
10	1.60	1.83	2.32
11	1.57	1.79	2.25
12	1.55	1.75	2.18
13	1.52	1.72	2.13
14	1.50	1.69	2.08
15	1.49	1.67	2.04
16	1.47	1.64	2.00
17	1.46	1.62	1.97
18	1.44	1.60	1.93
19	1.43	1.59	1.90
20	1.42	1.57	1.88
21	1.41	1.56	1.85
22	1.40	1.54	1.83
23	1.39	1.53	1.81
24	1.38	1.52	1.79
25	1.38	1.51	1.77
26	1.37	1.50	1.76
27	1.36	1.49	1.74
28	1.35	1.48	1.72
29	1.35	1.47	1.71
30	1.34	1.46	1.70

注：该表给出了 $F_{m,\infty}$ 分布的第 90、第 95 和第 99 个百分位数。它们可作为在 10%，5% 和 1% 的显著性水平下检验的临界值。



附表 5A　F_{n_1,n_2}分布的临界值——10%的显著性水平

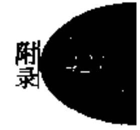
分母自由度	分子自由度(n_1)									
(n_2)	1	2	3	4	5	6	7	8	9	10
1	39.86	49.50	53.59	55.83	57.24	58.20	58.90	59.44	59.86	60.20
2	8.53	9.00	9.16	9.24	9.29	9.33	9.35	9.37	9.38	9.39
3	5.54	5.46	5.39	5.34	5.31	5.28	5.27	5.25	5.24	5.23
4	4.54	4.32	4.19	4.11	4.05	4.01	3.98	3.95	3.94	3.92
5	4.06	3.78	3.62	3.52	3.45	3.40	3.37	3.34	3.32	3.30
6	3.78	3.46	3.29	3.18	3.11	3.05	3.01	2.98	2.96	2.94
7	3.59	3.26	3.07	2.96	2.88	2.83	2.78	2.75	2.72	2.70
8	3.46	3.11	2.92	2.81	2.73	2.67	2.62	2.59	2.56	2.54
9	3.36	3.01	2.81	2.69	2.61	2.55	2.51	2.47	2.44	2.42
10	3.29	2.92	2.73	2.61	2.52	2.46	2.41	2.38	2.35	2.32
11	3.23	2.86	2.66	2.54	2.45	2.39	2.34	2.30	2.27	2.25
12	3.18	2.81	2.61	2.48	2.39	2.33	2.28	2.24	2.21	2.19
13	3.14	2.76	2.56	2.43	2.35	2.28	2.23	2.20	2.16	2.14
14	3.10	2.73	2.52	2.39	2.31	2.24	2.19	2.15	2.12	2.10
15	3.07	2.70	2.49	2.36	2.27	2.21	2.16	2.12	2.09	2.06
16	3.05	2.67	2.46	2.33	2.24	2.18	2.13	2.09	2.06	2.03
17	3.03	2.64	2.44	2.31	2.22	2.15	2.10	2.06	2.03	2.00
18	3.01	2.62	2.42	2.29	2.20	2.13	2.08	2.04	2.00	1.98
19	2.99	2.61	2.40	2.27	2.18	2.11	2.06	2.02	1.98	1.96
20	2.97	2.59	2.38	2.25	2.16	2.09	2.04	2.00	1.96	1.94
21	2.96	2.57	2.36	2.23	2.14	2.08	2.02	1.98	1.95	1.92
22	2.95	2.56	2.35	2.22	2.13	2.06	2.01	1.97	1.93	1.90
23	2.94	2.55	2.34	2.21	2.11	2.05	1.99	1.95	1.92	1.89
24	2.93	2.54	2.33	2.19	2.10	2.04	1.98	1.94	1.91	1.88
25	2.92	2.53	2.32	2.18	2.09	2.02	1.97	1.93	1.89	1.87
26	2.91	2.52	2.31	2.17	2.08	2.01	1.96	1.92	1.88	1.86
27	2.90	2.51	2.30	2.17	2.07	2.00	1.95	1.91	1.87	1.85
28	2.89	2.50	2.29	2.16	2.06	2.00	1.94	1.90	1.87	1.84
29	2.89	2.50	2.28	2.15	2.06	1.99	1.93	1.89	1.86	1.83
30	2.88	2.49	2.28	2.14	2.05	1.98	1.93	1.88	1.85	1.82
60	2.79	2.39	2.18	2.04	1.95	1.87	1.82	1.77	1.74	1.71
90	2.76	2.36	2.15	2.01	1.91	1.84	1.78	1.74	1.70	1.67
120	2.75	2.35	2.13	1.99	1.90	1.82	1.77	1.72	1.68	1.65

注:该表给出了F_{n_1,n_2}分布的第90个百分位数。它可作为在10%的显著性水平下检验的临界值。

附表 5B　F_{n_1,n_2}分布的临界值——5%的显著性水平

分母自由度(n_2)	分子自由度(n_1)									
	1	2	3	4	5	6	7	8	9	10
1	161.40	199.50	215.70	224.60	230.20	234.00	236.80	238.90	240.50	241.90
2	18.51	19.00	19.16	19.25	19.30	19.33	19.35	19.37	19.39	19.40
3	10.13	9.55	9.28	9.12	9.01	8.94	8.89	8.85	8.81	8.79
4	7.71	6.94	6.59	6.39	6.26	6.16	6.09	6.04	6.00	5.96
5	6.61	5.79	5.41	5.19	5.05	4.95	4.88	4.82	4.77	4.74
6	5.99	5.14	4.76	4.53	4.39	4.28	4.21	4.15	4.10	4.06
7	5.59	4.74	4.35	4.12	3.97	3.87	3.79	3.73	3.68	3.64
8	5.32	4.46	4.07	3.84	3.69	3.58	3.50	3.44	3.39	3.35
9	5.12	4.26	3.86	3.63	3.48	3.37	3.29	3.23	3.18	3.14
10	4.96	4.10	3.71	3.48	3.33	3.22	3.14	3.07	3.02	2.98
11	4.84	3.98	3.59	3.36	3.20	3.09	3.01	2.95	2.90	2.85
12	4.75	3.89	3.49	3.26	3.11	3.00	2.91	2.85	2.80	2.75
13	4.67	3.81	3.41	3.18	3.03	2.92	2.83	2.77	2.71	2.67
14	4.60	3.74	3.34	3.11	2.96	2.85	2.76	2.70	2.65	2.60
15	4.54	3.68	3.29	3.06	2.90	2.79	2.71	2.64	2.59	2.54
16	4.49	3.63	3.24	3.01	2.85	2.74	2.66	2.59	2.54	2.49
17	4.45	3.59	3.20	2.96	2.81	2.70	2.61	2.55	2.49	2.45
18	4.41	3.55	3.16	2.93	2.77	2.66	2.58	2.51	2.46	2.41
19	4.38	3.52	3.13	2.90	2.74	2.63	2.54	2.48	2.42	2.38
20	4.35	3.49	3.10	2.87	2.71	2.60	2.51	2.45	2.39	2.35
21	4.32	3.47	3.07	2.84	2.68	2.57	2.49	2.42	2.37	2.32
22	4.30	3.44	3.05	2.82	2.66	2.55	2.46	2.40	2.34	2.30
23	4.28	3.42	3.03	2.80	2.64	2.53	2.44	2.37	2.32	2.27
24	4.26	3.40	3.01	2.78	2.62	2.51	2.42	2.36	2.30	2.25
25	4.24	3.39	2.99	2.76	2.60	2.49	2.40	2.34	2.28	2.24
26	4.23	3.37	2.98	2.74	2.59	2.47	2.39	2.32	2.27	2.22
27	4.21	3.35	2.96	2.73	2.57	2.46	2.37	2.31	2.25	2.20
28	4.20	3.34	2.95	2.71	2.56	2.45	2.36	2.29	2.24	2.19
29	4.18	3.33	2.93	2.70	2.55	2.43	2.35	2.28	2.22	2.18
30	4.17	3.32	2.92	2.69	2.53	2.42	2.33	2.27	2.21	2.16
60	4.00	3.15	2.76	2.53	2.37	2.25	2.17	2.10	2.04	1.99
90	3.95	3.10	2.71	2.47	2.32	2.20	2.11	2.04	1.99	1.94
120	3.92	3.07	2.68	2.45	2.29	2.18	2.09	2.02	1.96	1.91

注：该表给出了F_{n_1,n_2}分布的第 95 个百分位数。它可作为在 5%的显著性水平下检验的临界值。



附表 5C　F_{n_1,n_2}分布的临界值——1%的显著性水平

分母自由度	分子自由度(n_1)									
(n_2)	1	2	3	4	5	6	7	8	9	10
1	4 052.00	4 999.00	5 403.00	5 624.00	5 763.00	5 859.00	5 928.00	5 981.00	6 022.00	6 055.00
2	98.50	99.00	99.17	99.25	99.30	99.33	99.36	99.37	99.39	99.40
3	34.12	30.82	29.46	28.71	28.24	27.91	27.67	27.49	27.35	27.23
4	21.20	18.00	16.69	15.98	15.52	15.21	14.98	14.80	14.66	14.55
5	16.26	13.27	12.06	11.39	10.97	10.67	10.46	10.29	10.16	10.05
6	13.75	10.92	9.78	9.15	8.75	8.47	8.26	8.10	7.98	7.87
7	12.25	9.55	8.45	7.85	7.46	7.19	6.99	6.84	6.72	6.62
8	11.26	8.65	7.59	7.01	6.63	6.37	6.18	6.03	5.91	5.81
9	10.56	8.02	6.99	6.42	6.06	5.80	5.61	5.47	5.35	5.26
10	10.04	7.56	6.55	5.99	5.64	5.39	5.20	5.06	4.94	4.85
11	9.65	7.21	6.22	5.67	5.32	5.07	4.89	4.74	4.63	4.54
12	9.33	6.93	5.95	5.41	5.06	4.82	4.64	4.50	4.39	4.30
13	9.07	6.70	5.74	5.21	4.86	4.62	4.44	4.30	4.19	4.10
14	8.86	6.51	5.56	5.04	4.69	4.46	4.28	4.14	4.03	3.94
15	8.68	6.36	5.42	4.89	4.56	4.32	4.14	4.00	3.89	3.80
16	8.53	6.23	5.29	4.77	4.44	4.20	4.03	3.89	3.78	3.69
17	8.40	6.11	5.18	4.67	4.34	4.10	3.93	3.79	3.68	3.59
18	8.29	6.01	5.09	4.58	4.25	4.01	3.84	3.71	3.60	3.51
19	8.18	5.93	5.01	4.50	4.14	3.94	3.77	3.63	3.52	3.43
20	8.10	5.85	4.94	4.43	4.10	3.87	3.70	3.56	3.46	3.37
21	8.02	5.78	4.87	4.37	4.04	3.81	3.64	3.51	3.40	3.31
22	7.95	5.72	4.82	4.31	3.99	3.76	3.59	3.45	3.35	3.26
23	7.88	5.66	4.76	4.26	3.94	3.71	3.54	3.41	3.30	3.21
24	7.82	5.61	4.72	4.22	3.90	3.67	3.50	3.36	3.26	3.17
25	7.77	5.57	4.68	4.18	3.85	3.63	3.46	3.32	3.22	3.13
26	7.72	5.53	4.64	4.14	3.82	3.59	3.42	3.29	3.18	3.09
27	7.68	5.49	4.60	4.11	3.78	3.56	3.39	3.26	3.15	3.06
28	7.64	5.45	4.57	4.07	3.75	3.53	3.36	3.23	3.12	3.03
29	7.60	5.42	4.54	4.04	3.73	3.50	3.33	3.20	3.09	3.00
30	7.56	5.39	4.51	4.02	3.70	3.47	3.30	3.17	3.07	2.98
60	7.08	4.98	4.13	3.65	3.34	3.12	2.95	2.82	2.72	2.63
90	6.93	4.85	4.01	3.53	3.23	3.01	2.84	2.72	2.61	2.52
120	6.85	4.79	3.95	3.48	3.17	2.96	2.79	2.66	2.56	2.47

注：该表给出了F_{n_1,n_2}分布的第99个百分位数。它可作为在1%的显著性水平下检验的临界值。



第2章

2.1　这些结果是随机的，因为在它们实际发生之前，是无法确切地知道它们的。你无法确切地知道你将要遇到的下一个人的性别、通勤到学校花费的时间等等。

2.2　如果 X 和 Y 是独立的，那么对于所有的 y 和 x 值，有 $\Pr(Y \leq y | X = x) = \Pr(Y \leq y)$。也就是说，独立性意味着 Y 的条件分布和边缘分布是相同的，所以，知道 X 的值不会改变 Y 的概率分布，知道 X 的值对 Y 取不同值的概率没有任何意义。

2.3　尽管降雨量和婴儿出生数之间没有明显的因果联系，但降雨量也能够告诉你一些关于婴儿出生数的事情。知道降雨量可以知道是什么季节，而出生是季节性的。因此，知道降雨量会告诉你关于月份的一些信息，这又会告诉你关于婴儿出生数的一些信息。所以，降雨量和婴儿出生数不是独立分布的。

2.4　4个随机选择的学生的平均体重不可能精确地等于145磅。4个学生的不同的组会有不同的样本平均体重，有的大于145磅，而有的小于145磅，因为这4个学生是随机选择的，他们的样本平均体重也是随机的。

2.5　所有这些分布都具有正态的形状，且以 Y 的均值1为中心，但它们会有不同的“离散度”，因为它们有不同的方差。$\bar{Y}$ 的方差是 $4/n$，所以随着 n 的增加，方差减小。在图形中，$n=2$ 时正态分布的离散度要比 $n=10$ 时宽，而 $n=10$ 时的离散度应该比 $n=100$ 时宽。当 n 变得非常大时，方差接近于0，正态分布退缩到 Y 的均值周围，即随着 n 的增大，$\bar{Y}$ 的分布高度集中在 μ_Y 周围（$\bar{Y}$ 逼近 μ_Y 的概率趋向于1），这正是大数定律所表明的。

2.6　当 $n=5$ 时，正态逼近的效果看上去不是很好，但当 $n=25$ 和 $n=100$ 时，正态逼近的效果看上去就会很好。因此，当 $n=25$ 或 $n=100$ 时，$\Pr(\bar{Y} \leq 0.1)$ 大约等于由正态逼近所计算的值，但当 $n=5$ 时却不能很好地被正态分布的值所近似。

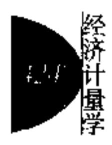
第3章

3.1　总体均值是总体中的平均数。样本平均数 $\bar{Y}$ 是从总体中抽取的一个样本的平均数。

3.2　估计量是计算总体参数值的一个理性猜测的方法，如总体均值。估计值是估计量在一个给定的样本中所生成的数值。$\bar{Y}$ 就是估计量的一个例子，它给出了计算总体均值的一个理性猜测的一种方法（将样本中所有的 Y 值相加然后除以 n）。如果一个 $n=4$ 的样本生成的样本 Y 的值为100，104，123和96，那么用估计量 $\bar{Y}$ 计算的估计值是105.75。

3.3　在所有的情况下 $\bar{Y}$ 的均值都是10。$\bar{Y}$ 的方差是 $\text{var}(Y)/n$，当 $n=10$ 时，$\text{var}(\bar{Y}) = 1.6$；当 $n=100$ 时，$\text{var}(\bar{Y}) = 0.16$；当 $n=1\ 000$ 时，$\text{var}(\bar{Y}) = 0.016$。因为随着 n 的增加，$\text{var}(\bar{Y})$ 收敛于0，所以随着 n 的增加，$\bar{Y}$ 逼近于10的概率接近于1。这是大数定律所表明的。

3.4　当用样本均值检验假设时，中心极限定理起着重要的作用。因为当样本容量很大时，样本均值近似为正态分布，所以，假设检验的临界值和检验统计量的 p 值可以利用正态分布来计算。正态临界值还被用于构造置信区间。



3.5　这部分内容在3.2节中做了描述。

3.6　置信区间包含了当被用做零假设时不能被拒绝的参数的所有值（如均值）。因此，它概括了从一个数量非常大的假设检验中所得的结果。

3.7　图(a)是向上倾斜的，而且点恰好在直线上。图(b)是向下倾斜的，而且点恰好在

直线上。图(c)应该显示出正向关系,而且点应该接近于但不是完全恰好在向上倾斜的直线上。图(d)显示出变量之间的一般的负向关系,而且点散布在向下倾斜的直线周围。图(e)在变量之间没有明显的线性关系。

第4章

4.1 β_1 是总体回归中斜率的值,它是未知的。$\hat{\beta}_1$(估计量)给出了从一个样本中估计未知值的 β_1 的表达式。同样,u_i 是第 i 个观测值的回归误差值;u_i 是 Y_i 与总体回归线 $\beta_0+\beta_1X_i$ 之差。因为 β_0 和 β_1 的值都是未知的,所以 u_i 的值也是未知的。相反,$\hat{u}_i$ 是 Y_i 与 $\hat{\beta}_0+\hat{\beta}_1X_i$ 之差,因而,$\hat{u}_i$ 是 u_i 的一个估计量。最后,$E(Y|X)=\beta_0+\beta_1X$ 是未知的,因为 β_0 和 β_1 是未知的;对它的一个估计量是普通最小二乘预测值 $\hat{\beta}_0+\hat{\beta}_1X$。

4.2 H_0: $\mu=0$ 的双边检验的 p 值,可以使用 $Y_i,i=1,\cdots,n$ 的独立同分布观测值的集合分三步来构造:(1)计算样本均值和标准误 $SE(\bar{Y})$;(2)计算这个样本的 t 统计量 $t^{act}=\bar{Y}^{act}/SE(\bar{Y})$;(3)利用标准正态分布表计算 p 值 $=\Pr(|Z|>|t^{act}|)=2\phi(-|t^{act}|)$。可以用类似的三步程序来构造双边检验 H_0: $\beta_1=0$ 的 p 值:(1)计算回归斜率的 OLS 估计值和标准误 $SE(\hat{\beta}_1)$;(2)计算这个样本的 t 统计量 $t^{act}=\hat{\beta}_1^{act}/SE(\hat{\beta}_1)$;(3)利用标准正态分布表计算 p 值 $=\Pr(|Z|>|t^{act}|)=2\phi(-|t^{act}|)$。

4.3 1992 年的工资性别差异可以用公式(4.41)中的回归和表 3—1 中 1992 年的行所总结的数据来估计。因变量是样本中第 i 个人的每小时收入。解释变量是一个二元变量,如果是男性,它等于1;如果是女性,它等于0。总体中的工资性别差异是回归中的总体系数 β_1,它可以用 $\hat{\beta}_1$ 来估计。其他年份的工资性别差异可以用类似的方式估计。

4.4 R^2 值表明了点在所估计的回归直线周围的分散程度。当 $R^2=0.9$ 时,点应该散布在非常接近回归直线的地方。当 $R^2=0.5$ 时,点在直线周围应该是更分散的。R^2 并不表明直线的斜率是正的还是负的。

第5章

5.1 由于存在遗漏变量,因此 $\hat{\beta}_1$ 可能是有偏的。富裕学区的学校可能在所有的教育投入上花费更多,进而会有更小的班级规模,在图书馆有更多的书和更多的计算机。这些其他投入可能会导致更高的平均考试成绩。因此,$\hat{\beta}_1$ 是向上偏的,因为每位学生的计算机数与对平均考试成绩有正效应的遗漏变量正相关。

5.2 如果 X_1 增加 3 个单位而 X_2 不变,那么预期 Y 的变化为 $3\beta_1$ 个单位。如果 X_2 减少 5 个单位而 X_1 不变,那么预期 Y 的变化为 $-5\beta_2$ 个单位。如果 X_1 增加 3 个单位而 X_2 减少 5 个单位,那么预期 Y 的变化为 $3\beta_1-5\beta_2$ 个单位。

5.3 该回归不能确定回归因子中一个回归因子变化(假设其他回归因子不变)的效应,因为如果一个完全多重共线的回归因子的值保持不变,那么其他回归因子的值也不变,即在一个多重共线的回归因子中不存在独立的变化。完全多重共线的回归因于的两个例子是:(1)用磅测量的某人的体重和用千克测量的同一个人的体重;(2)当数据取自于所有男生学校时,学生中男生的比重和常数项。

5.4 $\beta_1=0$ 的零假设可以用重要概念 5.4 中所介绍的 β_1 的 t 统计量进行检验。同样,$\beta_2=0$ 的零假设可以用 β_2 的 t 统计量进行检验。$\beta_1=0$ 和 $\beta_2=0$ 的零假设可以用 5.7 节中

的 F 统计量进行检验。F 统计量对检验一个联合假设是必需的，因为该检验将以 $\hat{\beta}_1$ 和 $\hat{\beta}_2$ 为基础，这意味着该检验的程序必须使用它们的联合分布的性质。

5.5　这里有一个例子。利用几年的经济计量学授课考试的数据，一位教授用学生的期末考试分数(Y)对他(她)们的期中考试分数(X)进行回归。这个回归将会有高的 R^2，因为期中考试成绩好的人倾向于在期末考试中也取得好成绩。然而，这个回归产生了期中考试分数对期末考试分数因果效应的一个有偏估计。期中成绩好的学生倾向于是那些上课有规律、学习努力而且对这个科目有热情的学生。这些变量与期中考试分数相关，但又是期末考试分数的决定性因素，因此漏掉它们会导致遗漏变量偏差。

第6章

6.1　这个回归函数看上去像图6—3中的二次回归或图6—4中的对数函数。前者设定为 Y 对 X 和 X^2 的回归，后者设定为 Y 对 $\ln(X)$ 的回归。有许多经济关系类似于这种形状。例如，这个形状可能代表生产函数中递减的边际劳动生产率。

6.2　对方程的两边取对数得：$\ln(Q)=\beta_0+\beta_1\ln(K)+\beta_2\ln(L)+\beta_3\ln(M)+u$，其中 $\beta_0=\ln(\lambda)$。该生产函数的参数可以通过用产量的对数对资本、劳动和原材料的对数进行回归来估计。

6.3　GDP 增加2%意味着 $\ln(GDP)$ 增加0.02。$\ln(m)$ 隐含的变化是 $1.0\times0.02=0.02$，它对应于 m 增加2%。由于 R 用百分点测量，因此，R 从4.0增加到5.0，或者说增加1.0个百分点。这导致 $\ln(m)$ 的变化为 $-0.02\times1.0=-0.02$，它对应于 m 下降2%。

6.4　你想要比较你所做的线性回归的拟合程度与非线性回归的拟合程度。答案将取决于你所选择的用来比较的非线性回归。你可以检验线性回归，对应于向线性回归中添加 X^2 得到的一个二次回归。如果 X^2 的系数显著地不同于0，那么就可以拒绝这个关系是线性的零假设，接受它是二次的备择假设。

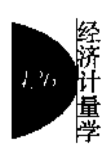
6.5　在问题6.2的方程中增加一个交叉项得到：$\ln(Q)=\beta_0+\beta_1\ln(K)+\beta_2\ln(L)+\beta_3\ln(M)+\beta_4[\ln(K)\times\ln(L)]+u$。$\ln(L)$ 对 $\ln(Q)$ 的局部效应现在是 $\beta_2+\beta_4\ln(K)$。

第7章

7.1　见重要概念7.1和紧接在重要概念方框之后的段落。

7.2　包含一个本来属于回归的额外变量将会消除或减少遗漏变量偏差。但是一般而言，包含一个不属于回归的额外变量将会降低其他系数估计量的精度(增加方差)。

7.3　区别 Y 中的测量误差或 X 中的测量误差是非常重要的。如果 Y 的测量中有错误，那么这个测量误差会成为回归误差项 u 的一部分。如果重要概念5.4中的假设仍然成立，那么这不会影响普通最小二乘回归的内部有效性，尽管使回归误差项的方差变大了，但它增加了普通最小二乘估计量的方差。可是，如果 X 的测量中有错误，那么这可能会导致该回归因子和回归误差项之间的相关，从而导致普通最小二乘估计量的不一致性。如公式(7.2)所表明的，随着这个不一致性变得越严重，这个测量误差就越大(即公式(7.2)中的 σ_w^2 越大)。



7.4　学生成绩较高的学校更可能会自愿参加这项考试，因此，自愿参加此项考试的学校不能代表学校的总体，并会导致样本选择偏差。例如，如果所有低学生—教师比的学校参加了这个考试，但在高学生—教师比的学校中只有成绩最好的学校参加了此项考试，那么所估计的班级规模效应将是有偏的。

7.5　城市的高犯罪率可能会决定他们需要更多的警察保护和在警察上的更多的花费，但如果警察尽职尽责，那么较多的警察经费支出会减少犯罪。因此，存在从犯罪率到警察经费支出和从警察经费支出到犯罪率的因果联系，这样就导致了联立因果偏差。

7.6　如果这个回归有同方差误差，那么同方差的和异方差的标准误一般是相近的，因为二者都是一致的。但是，如果误差项是异方差的，那么同方差标准误就是不一致的，而异方差标准误是一致的。因而，两个标准误的不同值构成了异方差的证据，这表明了应该使用异方差标准误。

第8章

8.1　面板数据（也称纵向数据）是指 n 个不同的实体在 T 个不同的时期观测到的数据。一个下角标 i 表示实体，而另一个下角标 t 表示时期。

8.2　一个人的能力或动机既可能影响其教育，也可能影响其收人。能力较强的个人倾向于完成较多年的教育，而且对给定的教育水平而言，他们倾向于有较高的收人。这同样也适用于有动机的人。宏观经济状况也是一个影响收人和教育的随时间而变化的变量。在经济衰退期间，失业率高，收入低，大学入学率增加。因人而异和因时间而异的固定效应可以被包含在回归中，以控制因人而异和因时间而异的变量。

8.3　当因人而异的固定效应被包含在回归模型中时，他们捕捉到了在样本时期不发生变化的个体的全部特征。由于性别在样本时期不发生变化，因此，它对收人的效应不能通过从因人而异的固定效应中分开来确定。同样，时间固定效应捕捉到了在个体之间不发生变化的时期的全部特征。全国失业率水平在给定的时点上对样本中所有的个体而言是相同的，因而它对收入的效应不能通过从因时间而异的固定效应中分开来确定。

第9章

9.1　由于 Y 是二元变量，因此它的预测值是 $Y=1$ 的概率。这个概率必须在 0 和 1 之间，所以概率值 1.3 是没有意义的。

9.2　第(1)列中的结果是针对线性概率模型的。线性概率模型中的系数表示 X 的单位变化对 $Y=1$ 的概率的影响。第(2)列和第(3)列中的结果是针对 logit 模型和 probit 模型的，这些系数很难解释。要计算 logit 模型和 probit 模型中 X 的单位变化对 $Y=1$ 的概率的影响，需使用重要概念 9.2 中提出的方法和步骤。

9.3　她应该使用 logit 模型或 probit 模型。这些模型要好于线性概率模型，因为它们将回归的预测值限制在 0 和 1 之间。通常，probit 和 logit 回归会给出相似的结果，因此，她应该使用她的软件易于执行的方法。

9.4　不能使用 OLS，因为该回归函数不是回归系数的线性函数（这些系数出现在非线性函数 ϕ 或 F 的内部）。极大似然估计量是有效的，能够处理参数是非线性的回归函数。

第10章

10.1　回归误差 u 的增加使需求曲线向外移动，导致价格和数量都增加，因而 $\ln(P^{butter})$ 与误差项是正相关的。由于这个正的相关性，β_1 的 OLS 估计量是不一致的，并且可能比 β_1 的真实值大。

10.2　这个州的人均树木数量是外生的，因为它与需求函数中的误差项可以被认为是不相关的。然而，因为它也可能与 $\ln(P^{cigarettes})$ 不相关，所以它并不是相关的。一个有效的



工具变量必须是外生的且是相关的，所以这个州的人均树木数量不是一个有效的工具变量。

10.3　可以论证，律师的数量与监禁率是相关的，所以它是相关的（尽管这应该用 10.3 节中的方法来检查）。然而，因为具有高于预期犯罪率的州（具有正的回归误差）可能有较多的律师（罪犯必须被辩护或起诉），所以，律师的数量将与回归误差正相关。这意味着律师的数量不是外生的。一个有效的工具变量必须是外生的且是相关的，所以律师的数量不是一个有效的工具变量。

10.4　如果距离的差是一个有效的工具变量，那么它一定与 X 相关。在这个例子中，X 是一个表示病人是否接受心脏导管插入术的二元变量。工具变量的相关性可以用 10.3 节中总结的方法和步骤来检验。检查工具变量的外生性更为困难。如果存在比内生回归因子更多的工具变量，那么工具变量的联合外生性可用重要概念 10.6 中总结的 J 检验来检验。但是，如果工具变量的数量等于内生回归因子的数量，那么不可能从统计上检验外生性。在 McClellan，McNeil 和 Newhouse 的研究（1994）中，存在一个内生回归因子（治疗）和一个工具变量（距离之差），所以不能使用 J 检验。评估这个外生性需要使用专家判断。

第 11 章

11.1　最好随机地将处理水平分配到每个地块中。该问题中所列出的研究计划可能有缺陷，因为不同的地块组可能系统地不同。例如，前 25 块地可能比其他地块有更差的排水设备，这会导致较低的作物产量。问题中所列出的处理分配将这 25 块地放在控制组中，因而过度地估计了施肥对农作物产量的因果效应，这个问题可以用处理的随机分配来避免。

11.2　处理效应可被估计为处理组和非处理组（控制组）的平均胆固醇水平之差。利用公式（11.2）中所显示的具有额外回归因子的差分估计量，每位病人的体重、年龄和性别的数据可被用来改进这个估计值。这个回归可能会生成一个更精确的估计值，因为它控制了这些可能会影响胆固醇的额外因素。如果你有在进入实验之前每位病人的胆固醇水平数据，那么可以使用差分再差分估计量。这个估计量控制了那些在样本期间保持不变的胆固醇水平的因个体而异的决定性因素，如该人的导致高胆固醇的遗传因素。

11.3　如果转到小班的学生系统地不同于其他的学生，那么内部有效性就被打折扣了。例如，如果转学的学生倾向于有较高的收入，并且在校外有较多的学习机会，那么他们倾向于在标准化考试中表现得更好。该实验会错误地将这个成绩归功于较小的班级规模。最初，随机分配的信息可以像公式（11.6）那样被用做回归中的工具变量，来恢复内部有效性。初始的随机分配是个有效的工具变量，因为它是外生的（与回归误差项无关），并且是相关的（与实际分配相关）。

11.4　Hawthorne 效应不可能是施肥例子中的问题，除非（例如）工人是否认真地耕种不同的地块取决于处理。胆固醇实验中的病人可能比不在该实验中的病人更勤于吃药，使得胆固醇实验成为双方不知情的，所以医生和病人都不知道病人是在接受处理还是在接受安慰剂，这会降低实验效应。在像 STAR 这样的实验中，如果教师感觉到该实验给他们提供了一个证明小班是最好的这样的机会，那么实验效应可能是重要的。

11.5　地震在班级规模中引入了随机性，这使得处理看上去仿佛是被随机分配的。10.1 节中的讨论描述了工具变量回归如何利用这个引致的班级规模的变化来估计班级规模对考试分数的效应。

$\beta_1 FDD_t + \beta_2 FDD_{t-1} + \beta_3 FDD_{t-2} + \cdots + \beta_7 FDD_{t-6} + E(u_t | FDD_{t+1}, FDD_t, FDD_{t-1}, \cdots)$。当 FDD 是严外生的时，$E(u_t | FDD_{t+1}, FDD_t, FDD_{t-1}, \cdots) = 0$，因此，$FDD_{t+1}$ 不进入这个回归方程中。当 FDD_t 是外生的但不是严外生的时，它可能是 $E(u_t | FDD_{t+1}, FDD_t, FDD_{t-1}, \cdots) \neq 0$ 的情况，因此，FDD_{t+1} 将会进入到该回归方程中。

第 14 章

14.1　这位宏观经济学家想要构造 9 个变量的预测。如果在 VAR 中使用每个变量的四阶滞后，那么每个 VAR 方程中将包含 37 个回归系数（常数项和 9 个变量中每个变量 4 个系数）。样本期间包含 128 个季度观测值。当用这 128 个观测值估计 37 个系数时，所估计的系数可能是不精确的，这导致不准确的预测。一种可供选择的方法是对每个变量使用单变量自回归。这个方法的优势是可以估计相对较少的参数，因此，系数将会被 OLS 精确地估计；缺陷是该预测只利用了被预测变量的滞后值，而其他变量的滞后值也可能包含额外的有用的预测信息。一种折中方法是，使用一组含额外预测因子的时间序列回归，例如，可能使用 GDP、消费和长期利率的滞后设定 GDP 的预测回归方程，但排除了其他的变量。可以使用短期利率、长期利率、GDP 和通货膨胀率的滞后设定短期利率预测回归方程，基本思路是，在每个回归方程中包含最重要的预测因子，省略不是非常重要的变量。

14.2　Y_{t+2} 的预测值是 $Y_{t+2|t} = 0.7^2 \times 5 = 2.45$，$Y_{t+30}$ 的预测值是 $Y_{t+30|t} = 0.7^{30} \times 5 = 0.0001$。结果是合理的，因为该过程是适度序列相关的（$\beta_1 = 0.7$），那么 Y_{t+30} 与 Y_t 只是弱相关的，这意味着 Y_{t+30} 的预测值应该非常逼近于 μ_Y，即 Y 的均值。既然该过程是平稳的，并且 $\beta_0 = 0$，那么 $\mu_Y = 0$。因此，就像预期的一样，$Y_{t+30|t}$ 非常逼近于 0。

14.3　如果 Y 和 C 是协整的，那么误差修正项 $Y-C$ 也是平稳的。序列 $Y-C$ 的图形应该表现出平稳性。序列 $Y-C$ 的协整检验可以使用迪基—富勒或 DF-GLS 单位根检验来完成。这是协整系数为已知条件下检验协整的一个例子。

14.4　当 u_{t-1}^2 异常地大时，σ_t^2 也会很大。由于 σ_t^2 是 u_t 的条件方差，因此，u_t^2 可能也很大，这将会导致一个大的 σ_{t+1}^2 值，依此类推。

14.5　当零假设是假的时，一个功效强大的检验更有可能拒绝该零假设。这提高了你区分单位 AR 根和小于 1 的根的能力。

第 15 章

15.1　如果重要概念 15.1 中的假设 4 是真的，那么在大样本条件下，用异方差稳健的标准误构造的一个 95% 的置信区间将会以 95% 的概率包含 β_1 的真实值。如果重要概念 15.1 中的假设 4 是假的，那么同方差惟一的方差估计量是不一致的。因此一般而言，在大样本条件下，如果误差是异方差的，那么用同方差惟一的标准误构造的一个 95% 的置信区间不会以 95% 的概率包含 β_1 的真实值，所以这个置信区间不是渐近有效的。

15.2　根据 Slutsky 定理，$A_n B_n$ 服从渐近的 $N(0,9)$ 分布。因此，$\Pr(A_n B_n < 2)$ 约等于 $\Pr[Z < (2/3)]$，这里 Z 是标准正态随机变量。计算这个概率得到 $\Pr[Z < (2/3)] = 0.75$。



15.3　对于 $X_t \leq 10$ 的值，这些点应该非常接近于该条回归直线，因为 u_t 的方差很小。当 $X_t > 10$ 时，这些点应该离该条回归直线非常远，因为 u_t 的方差很大。由于 $X_t \leq 10$ 的这些点更接近于回归直线，因此 WLS 给予它们更大的权重。

15.4　高斯—马尔可夫定理隐含着被平均的估计量不可能比 WLS 好。为了理解这一点，注意被平均的估计量是 $Y_1, \cdots, Y_n$ 的一个线性函数（OLS 估计量是线性函数，因此它们的

均值也是线性函数)并且是无偏的(OLS 估计量是无偏的,因此它们的均值也是无偏的)。高斯—马尔可夫定理隐含着 WLS 是最佳的线性条件无偏估计量,因此,被平均的估计量不可能比 WLS 好。

第 16 章

16.1 $\boldsymbol{X}$ 的第一列的每一个元素都为 1,第二列和第三列中的元素都是 0 和 1。矩阵 $\boldsymbol{X}$ 的第一列是第二列与第三列的和,因此这些列是线性依赖关系,$\boldsymbol{X}$ 不具有满秩。通过剔除 X_{1i} 或 X_{2i} 来重新设定这个回归。

16.2 a. 用 OLS 估计这些回归系数,并计算异方差稳健的标准误。将置信区间构造为 $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$。

b. 用 OLS 估计这些回归系数,并计算异方差稳健的标准误。将置信区间构造为 $\hat{\beta}_1 \pm 1.96SE(\hat{\beta}_1)$。另一种可供选择的方法是,计算同方差惟一的标准误 $\widetilde{SE}(\hat{\beta}_1)$,将置信区间构造为 $\hat{\beta}_1 \pm 1.96\ \widetilde{SE}(\hat{\beta}_1)$。

c. 该置信区间可以按(b)的方式来构造,它们用到了大样本正态逼近。在假设 1 ~6 都为真的条件下,这个精确的分布可被用来构造置信区间 $\hat{\beta}_1 \pm t_{n-k-1,0.975}\widetilde{SE}(\hat{\beta}_1)$,其中 $t_{n-k-1,0.975}$ 是自由度为 $n-k-1$ 的 t 分布的第 97.5 个百分位数,这里 $n=500, k=1$。附表 2 的一个扩展形式显示了 $t_{498,0.975}=1.9648$。

16.3 不成立,这个结果要求有正态分布误差。

16.4 这里的 BLUE 估计量就是 GLS 估计量。你必须知道 $\boldsymbol{\Omega}$ 以计算精确的 GLS 估计量。然而,如果 $\boldsymbol{\Omega}$ 是某些参数的一个已知函数,而这些参数又可以被一致地估计出来,那么这些参数的估计量可被用来构造协方差矩阵 $\boldsymbol{\Omega}$ 的一个估计量。于是,这个估计量可被用来构造 GLS 估计量的一个可行形式。当样本容量很大时,这个估计量约等于 BLUE 估计量。

16.5 有许多例子,这只是其中的一个。假设 $X_t=Y_{t-1}$,u_t 是均值为 0 且方差为 σ^2 独立同分布的(也就是说,这个回归模型是第 12 章中的 AR(1) 模型)。在这种情况下,对于 $j<i$,X_i 依赖于 u_j,但对于 $j\geq i$,X_i 则不依赖于 u_j,这意味着 $E(u_i|X_i)=0$。然而,$E(u_{i-1}|X_i)\neq 0$,这意味着 $E(\boldsymbol{U}|\boldsymbol{X})\neq \boldsymbol{0}_n$。

autoregressive distributed lag model 自回归分布滞后模型：将某一时间序列变量 Y_t 表示为 Y_t 滞后值和另一变量 X_t 滞后值的函数的一个线性回归模型。该模型表示为 ADL(p,q)，其中 p 表示 Y_t 的滞后期数，q 表示 X_t 的滞后期数。

average causal effect 平均因果效应：在一个异质总体中，个体因果效应的总体平均值，也称为平均处理效应。

balanced panel 均衡面板：一个没有丢失观测值的面板数据集。在这个数据集中，变量在每个实体和每个时期都被观测到。

base specification 基准设定：使用专家判断、经济理论以及数据是如何被搜集上来的知识的组合，选择一组回归因子的基线回归设定或基准回归设定。

Bayes information criterion 贝叶斯信息准则：请见信息准则。

Bernoulli distribution 贝努里分布：一个贝努里随机变量的概率分布。

Bernoulli random variable 贝努里随机变量：只取 0 和 1 两个值的随机变量。

best linear unbiased estimator 最佳线性无偏估计量：指所有那些与样本值 Y 为线性函数关系且是无偏的估计量中方差最小的那个估计量。在高斯—马尔可夫条件下，OLS 估计量就是回归系数的最佳线性无偏估计量。

bias 偏差：一个估计量与所要估计的参数值之差的期望值。如果 $\hat{\mu}_Y$ 是 μ_Y 的一个估计量，那么 $\hat{\mu}_Y$ 的偏差就是 $E(\hat{\mu}_Y)-\mu_Y$。

BIC：请见信息准则。

binary variable 二元变量：一个或者取 1 或者取 0 的变量。一个二元变量被用来表示一个二元的结果。例如，对于一个人的性别来说，如果该人是女性，那么 $X=1$；如果该人是男性，那么 $X=0$。因此，X 就是反映个人性别的一个二元（或指示，或虚拟）变量。

bivariate normal distribution 二元正态分布：描述两个随机变量联合分布的正态分布的一个推广形式。

BLUE：请见最佳线性无偏估计量。

break date 突变日期：指总体时间序列回归系数的一个离散变化的日期。

causal effect 因果效应：在一个理想化随机控制实验中所测度的一个给定干预或处理所产生的期望效应。

central limit theorem 中心极限定理：数理统计中的一个结果。该定理指出，在一般条件下，当样本规模很大时，标准化的样本平均数的抽样分布可被一个标准正态分布很好地逼近。

chi-squared distribution 卡方分布（或 χ^2 分布）：m 个独立的标准正态随机变量平方和的分布。参数 m 被称为卡方分布的自由度。

Chow test 邹检验：已知突变日期时，对时间序列回归中的突变进行检验的一种方法。

coefficient of determination 判定系数：请见 R^2。

cointegration 协整：指两个或两个以上时间序列变量拥有一个共同的随机趋势。

common trend 共同趋势：由两个或更多个时间序列变量所共同拥有的趋势。

conditional distribution 条件分布：一个随机变量在已知另一个随机变量取某一特定值条件下的概率分布。



conditional expectation 条件期望：一个随机变量在已知另一个随机变量取某一特定值条件下的期望值。

conditional heteroskedasticity 条件异方差：通常指依赖于其他变量的一个误差项的方差。

conditional mean 条件均值：一个条件分布的均值。请见条件期望。

conditional mean independence 条件均值独立性：给定回归因子条件下，回归误差项 u_i 的条件期望依赖于回归因子中的一部分而不是全部。

conditional variance 条件方差：即某一条件分布的方差。

confidence interval (or confidence set) 置信区间（或置信集）：按事先设定的概率在重复样本间计算时，包含总体参数真实值的一个区间（或集合）。

confidence level 置信水平：一个置信区间（或置信集）包含所估计参数真实值的一个事先设定的概率。

consistency 一致性：指一个估计量是一致的。请见一致估计量。

consistent estimator 一致估计量：依概率收敛到所估计的参数的估计量。

continuous random variable 连续随机变量：一个能取连续值的随机变量。

control group 控制组：在一项实验中，没有接受处理或干预的组。

control variable 控制变量：回归因子的另外一个术语，具体地讲，指能够控制那些决定因变量变化的因素之一的回归因子。

convergence in distribution 依分布收敛：当一个分布序列收敛到一个极限时。15.2 节中给出了精确的定义。

convergence in probability 依概率收敛：当一个随机变量序列收敛到某一特定的值时。例如，随着样本规模的增大，样本平均值变得越来越趋近于总体平均值。请见重要概念 2.6 和 15.2 节。

correlation 相关性：两个随机变量一起移动或变化的程度的一个无量纲测量指标。X 和 Y 之间的相关性（或相关系数）是 $\sigma_{XY}/\sigma_X\sigma_Y$，表示为 $corr(X,Y)$。

correlation coefficient 相关系数：请见相关性。

covariance 协方差：两个随机变量一起变化的程度的一个测量指标。X 和 Y 之间的协方差是期望值 $E[(X-\mu_X)(Y-\mu_Y)]$，用 $\text{cov}(X,Y)$ 或 σ_{XY} 表示。

covariance matrix 协方差矩阵：由随机变量向量的方差和协方差构成的矩阵。

critical value 临界值：一个检验统计量的值，在这个值上，给定显著性水平下该检验正好拒绝了零假设。

cross-sectional data 截面数据：所搜集的不同实体在单个时期的数据。

cubic regression model 三次回归模型：将 X, X^2 和 X^3 作为回归因子的一个非线性回归函数。

cumulative distribution function (c.d.f.) 累积分布函数：请见累积概率分布。

cumulative dynamic multiplier 累积动态乘数（或累积动态乘子，译者注）：时间序列变量 X 的单位变化对 Y 的累积效应。h 期累积动态乘数就是 X_t 的单位变化对 $Y_t+Y_{t+1}+\cdots+Y_{t+h}$ 的效应。

cumulative probability distribution 累积概率分布：反映一个随机变量小于或等于某一给定值这个概率的一个函数。



dependent variable 因变量：在回归模型或其他统计模型中待解释的变量。该变量出现在回归模型的左边。

deterministic trend 确定性趋势：一个变量随时间的推移所表现出的一种持续的、长期的运动，它可被表示为一个时间的非随机函数。

Dickey-Fuller test 迪基—富勒检验：在一阶自回归（AR(1)）中检验单位根的一种方法。

difference estimator 差分估计量：构造为处理组和控制组的样本平均结果之差的一个因果效应估计量。

difference-in-difference estimator 差分再差分估计量：在处理组中 Y 的平均变化减去控制组中 Y 的平均变化。

discrete random variable 离散随机变量：一个取离散值的随机变量。

distributed lag model 分布滞后模型：回归因子是 X 的当前值和滞后值的一种回归模型。

dummy variable 虚拟变量（或哑变量，译者注）：请见二元变量。

dynamic causal effect 动态因果效应：一个变量对另一个变量的当前值和未来值的因果效应。

dynamic multiplier 动态乘数（或动态乘子，译者注）：h 期动态乘数就是时间序列变量 X_t 的单位变化对 Y_{t+h} 的效应。

endogenous variable 内生变量：一个与误差项相关的变量。

errors-in-variable bias 变量误差偏差：在回归因子中因测量误差所引起的一个回归系数估计量的偏差。

error term 误差项：Y 和总体回归函数之差，在本书中用 u 表示。

estimate 估计值：根据一个特定样本中的数据所计算的一个估计量的数值。

estimator 估计量：需要从一个总体中随机抽取的一个样本数据的函数。一个估计量就是使用样本数据计算总体参数值（如总体均值）的一个合理猜测值的步骤。

exact distribution 精确分布：一个随机变量的精确概率分布。

exact identification 恰好识别：当工具变量数等于内生回归因子数时。

exogenous variable 外生变量：一个与回归误差项不相关的变量。

expected value 期望值：一个随机变量在许多次重复的实验中或事件中发生的长期平均值。它是该随机变量所具有的所有可能取值的概率加权平均。Y 的期望值用 $E(Y)$ 表示，也称为 Y 的期望。

experimental data 实验数据：从一项实验中获得的数据，该项实验可能是用来评估一项处理，或评估一项政策，或调查一个因果效应。

experimental effect 实验效应：当实验的主体由于他们是实验的一部分而改变他们的行为时。

explained sum of squares (ESS) 解释的平方和：Y_i 的预测值 $\hat{Y}_i$ 与其平均值离差的平方和。请见方程(4.35)。

explanatory variable 解释变量：请见回归因子。

external validity 外部有效性：如果从一项统计研究的总体和环境中所得出的推断和结论，能够被推广到其他的总体和环境中，那么该统计研究的推断和结论就是外部有效的。

F-statistic F 统计量：用来对两个或两个以上回归系数的联合假设进行检验的统计量。

$F_{m,\infty}$-distribution $F_{m,\infty}$ 分布：具有 m 个自由度的卡方分布被 m 来除这样一个随机变量的分布。

feasible GLS 可行的 GLS：它是广义最小二乘（GLS）估计量的一种形式，它使用了不同观测值的回归残差之间条件协方差的一个估计量。

feasible WLS 可行的 WLS：加权最小二乘法（WLS）估计量的一种形式。它使用了回归误差项的条件方差的一个估计量。

first difference 一阶差分：一个时间序列变量 Y_t 的一阶差分是 $Y_t - Y_{t-1}$，用 ΔY_t 表示。



included endogenous variables 内含内生变量：与误差项相关的回归因子（通常在工具变量回归中）。

included exogenous variables 内含外生变量：与误差项不相关的回归因子（通常在工具变量回归中）。

independence 独立性：当知道一个随机变量的值不能提供关于另一个随机变量的值的信息时，如果两个随机变量的联合分布是它们的边缘分布的积，那么这两个随机变量是独立的。

indicator variable 指示变量：请见二元变量。

information criterion 信息准则：在一个自回归或一个分布滞后模型中，用来估计在上述回归中应该包含的滞后变量的个数的统计量。主要的例子是赤池信息准则（AIC）和贝叶斯信息准则（BIC）。

instrument 工具：请见工具变量。

instrumental variable 工具变量：一个与内生回归因子相关（工具变量相关性）但与回归误差项不相关（工具变量外生性）的变量。

instrumental variable (IV) regression 工具变量回归：当回归因子 X 与误差项 u 相关时，获得总体回归函数中未知系数的一致性估计量的一种方法。

interaction term 交互作用项：由另外两个回归因子的积所构成的一个回归因子，例如，$X_{1i} \times X_{2i}$。

intercept 截距：线性回归模型中 β_0 的值。

internal validity 内部有效性：在一项统计研究中，当关于因果效应的推断对所研究的总体有效时。

joint hypothesis 联合假设：由两个或两个以上单个假设所构成的假设，也就是说，对模型的参数有一个以上的约束。

joint probability distribution 联合概率分布：决定两个或两个以上随机变量结果的概率的概率分布。

lags 滞后值：一个时间序列变量在前一时期的值。Y_t 的第 j 个滞后值是 Y_{t-j}。

law of iterated expectations 累期望法则：概率论中的一个结论，即 Y 的期望值等于给定 X 条件下 Y 的条件期望的期望值，即 $E(Y)=E[E(Y|X)]$。

law of large numbers 大数定律：根据概率论中的这个结论，在一般条件下，当样本规模很大时样本平均数将以很高的概率接近于总体均值。

least squares assumptions 最小二乘假设：重要概念 4.3（单个变量回归）和重要概念 5.4（多元回归模型）中所列出的线性回归模型假设。

least squares estimator 最小二乘估计量：使残差平方和最小的那个估计量。

limited dependent variable 受限因变量：只能取有限值的因变量。例如，变量可能是只取 0～1 的二元变量，或附录 9.3 中所描述的任何一个模型中出现的变量。

linear-log model 线性对数模型：一个非线性回归函数，在这个函数中，因变量是 Y，自变量是 $\ln(X)$。



linear regression function 线性回归函数：具有常数斜率的回归函数。

linear probability model 线性概率模型：Y 是二元变量的回归模型。

logarithm 对数：为正的变量所定义的一个数学函数。它的斜率总是正的，但是趋向于 0。自然对数是指数函数的反函数，即 $X=\ln(e^X)$。

logit regression logit 回归(或称对数单位回归,译者注):一个二元因变量的非线性回归模型,其中总体回归函数是使用累积 logistic 分布函数建立的。

log-linear model 对数线性模型:一个非线性回归函数,其中因变量是 $\ln(Y)$,自变量是 X。

log-log model 双对数模型:一个非线性回归函数,其中因变量是 $\ln(Y)$,自变量是 $\ln(X)$。

longitudinal data 纵向数据:请见面板数据。

long-run cumulative dynamic multiplier 长期累积动态乘数(或长期累积动态乘子,译者注):时间序列变量 X 的变化对 Y 的累积长期效应。

marginal probability distribution 边缘概率分布:一个随机变量 Y 的概率分布的另一个名称,它将 Y 的单独的分布(边缘分布)同 Y 和另一随机变量的联合分布区别开来。

maximum likelihood estimator (MLE) 极大似然估计量:通过使似然函数最大化所得到的未知参数的一个估计量。请见附录 9.2。

mean 均值:一个随机变量的期望值。Y 的均值用 μ_Y 表示。

moments of a distribution 分布的矩:一个随机变量不同次幂的期望值。随机变量 Y 第 r 阶矩是 $E(Y^r)$。

multicollinearity 多重共线性:请见完全多重共线性和不完全多重共线性。

multiple regression model 多元回归模型:单个变量回归模型的一个扩展,它允许 Y 依赖于 k 个回归因子。

natural experiment 自然实验:请见准实验。

natural logarithm 自然对数:请见对数。

95% confidence set 95% 的置信集:具有 95% 的置信水平的置信集。请见置信区间。

nonlinear least squares 非线性最小二乘:当回归函数是未知参数的一个非线性函数时所适用的类似于 OLS 的一种方法。

nonlinear regression function 非线性回归函数:斜率不是常数的回归函数。

nonstationary 非平稳的:当一个时间序列变量及其滞后值的联合分布随时间的变化而变化时。

normal distribution 正态分布:连续随机变量最常用的一种类似于钟形的分布。

null hypothesis 零假设:在假设检验中被检验的假设,通常用 H_0 表示。

observational data 观测数据:它是基于对一项实验环境之外的实际行为进行观测或测量的数据。

observation number 观测期数:在一个数据集中,分配给每个实体的惟一的标识符号。

OLS estimator OLS 估计量:请见普通最小二乘估计量。

OLS regression line OLS 回归线:用 OLS 估计量代替总体系数的回归直线。

OLS residual OLS 残差:Y_i 和 OLS 回归线之差,本书中用 $\hat{u}_i$ 表示。

omitted variables bias 遗漏变量偏差:一个变量既是 Y 的一个决定性因素,又与某一回归因子相关,但这样的变量在回归中被遗漏了,从而所产生的估计量的偏差。



one-sided alternative hypothesis 单边备择假设:所研究的参数在零假设所给出的值的一边。

order of integration 单整阶数:为了使一个时间序列变量变得平稳,必须对该时间序列变量进行差分的次数。一个 p 阶单整的时间序列变量必须进行 p 次差分,用 $I(p)$ 表示。

ordinary least squares estimator 普通最小二乘估计量：使残差平方和最小的回归截距和斜率的估计量。

overidentification 过度识别：当工具变量数超过内含内生回归因子数时。

p-value *p* 值：假设零假设是正确的，抽取一个至少与实际所计算的统计量一样对零假设是不利的统计量的概率。*p* 值也被称为边缘显著性概率，是零假设可能被拒绝的最小的显著性水平。

panel data 面板数据：多个实体在多个时期被观测的数据。

parameter 参数：决定一个概率分布或一个总体回归函数特性的常数。

partial compliance 偏依赖性：在一项随机化实验中，当一些参与者没有遵守处理协议时就发生了偏依赖性。

partial effect 局部效应：在保持其他回归因子不变的条件下，变化回归因子中的一个回归因子对 Y 的效应。

perfect multicollinearity 完全多重共线性：当一个回归因子是其他回归因子的线性函数时，就发生了完全多重共线性。

polynomial regression model 多项式回归模型：将 $X, X^2, \cdots, X^r$ 作为回归因子的一个非线性回归函数，其中 r 是整数。

population 总体：所研究的全部实体，如人、公司或学区等。

population coefficients 总体系数：请见总体截距和斜率。

population intercept and slope 总体截距和斜率：在单变量的回归中 β_0（截距）和 β_1（斜率）的真实值或总体值。在多元回归中，有多个斜率系数（$\beta_1, \beta_2, \cdots, \beta_k$），每个斜率系数对应着一个回归因子。

population multiple regression model 总体多元回归模型：重要概念 5.2 中的多元回归模型。

population regression line 总体回归线：在一个单变量的回归中，总体回归线是 $\beta_0 + \beta_1 X_i$；在一个多元回归中，总体回归线是 $\beta_0 + \beta_1 X_{1i} + \beta_2 X_{2i} + \cdots + \beta_k X_{ki}$。

power 功效：当备择假设为真时，一个检验正确地拒绝零假设的概率。

predicted value 预测值：由 OLS 回归线所预测的 Y_i 的值，本书中用 $\hat{Y}_i$ 表示。

price elasticity 价格弹性：价格每增加 1% 所引起的需求量的百分比变化。

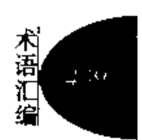
probability 概率：一个结果或一个事件从长期来看将要发生的次数的比重。

probability density function (p. d. f.) 概率密度函数：对一个连续随机变量而言，概率密度函数下任何两点间的面积就是该随机变量落在这两点间的概率。

probability distribution 概率分布：对一个离散型随机变量而言，该随机变量所能取的所有值的列表以及每个值所对应的概率。

probit regression Probit 回归（或称概率单位回归，译者注）：它是二元因变量的一个非线性回归模型，在这个模型中，使用累积标准正态分布函数对总体回归函数进行建模。

program evaluation 项目评估：它是涉及估计一个项目、一项政策、某些其他的干预或"处理"的效应的一个研究领域。



pseudo out-of-sample forecast 伪样本外预测：它是对部分样本数据值进行计算的预测，它使用看上去这些样本数据仿佛还没有实现的方法。

quadratic regression model 二次回归模型：以 X 和 X^2 作为回归因子的一个非线性回归函数。

quasi-experiment 准实验：它是指随机性是由个体环境的变化所引入的一种实验环境，因此看上去处理仿佛是被随机分配的。

R^2 拟合优度：在一个回归中，因变量的样本方差被回归因子所解释的部分。

$\overline{R}^2$ 调整的拟合优度：请见调整的 R^2 。

randomized controlled experiment 随机化控制实验：它是指这样一种实验，实验中参与者被随机地分配到没有接受处理的控制组，或被随机地分配到接受处理的处理组中。

random walk 随机游动：它是一个时间序列过程，在这个过程中，变量的值等于它的前期值加上一个不可预测的误差项。

random walk with drift 带漂移项的随机游动：随机游动的一个通式，在这个通式中，变量中的变化具有一个非零均值，但却是不可预测的。

regressand 回归方程中的从属变量：请见因变量。

regression specification 回归设定：对回归的一种描述，包括对所使用的回归因子及任何非线性变换的说明。

regressor 回归因子：出现在回归方程右边的变量，也即回归中的自变量。

rejection region 拒绝域：一个检验统计量的值的集合，在这个集合中该检验拒绝零假设。

repeated cross-sectional data 重复的截面数据：截面数据集的一个汇总，其中，每个截面数据集对应着一个不同的时期。

restricted regression 有约束的回归：为满足某些条件，回归系数被施加了约束的一种回归。例如，当计算经验规则 F 统计量时，这就是含有约束系数以满足零假设的回归。

root mean squared forecast error 均方根预测误差：预测误差平方平均值的平方根。

rule-of-thumb F-statistic 经验规则 F 统计量：在有约束和没有约束的回归中，使用残差平方和计算的 F 统计量。当回归误差是同方差的时，适合使用经验规则 F 统计量。

sample correlation 样本相关性：两个随机变量之间相关性的一个估计量。

sample covariance 样本协方差：两个随机变量之间协方差的一个估计量。

sample selection bias 样本选择偏差：当一个选择过程影响数据的可获得性，而且该过程与因变量相关时，所出现的回归系数估计量的偏差。

sample standard deviation 样本标准差：随机变量标准差的一个估计量。

sample variance 样本方差：随机变量方差的一个估计量。

sampling distribution 抽样分布：一个统计量在所有可能样本中的分布，即从同一总体中使用一个随机抽取的样本序列重复地评价该统计量所产生的分布。

scatterplot 散点图：由 X_i 和 Y_i 的 n 个观测值所组成的图形，其中每个观测值用点 (X_i, Y_i) 表示。

serial correlation 序列相关性：请见自相关。

serially uncorrelated 序列不相关的：所有自相关系数都等于零的时间序列变量。

significance level 显著性水平：在一个统计假设检验中，一个事先设定的当零假设为真时的拒绝概率。

simple random sampling 简单随机抽样：确保每个实体等可能地被选中，用这种方法从总体中随机地选择实体。

simultaneous causality bias 联立因果偏差：除了从 X 到 Y 有我们所感兴趣的因果关系外，从 Y 到 X 也存在相应的因果关系。联立因果关系使 X 与所研究总体回归中的误差项相关。

simultaneous equation bias 联立方程偏差：请见联立因果偏差。

unbalanced panel 非均衡面板：缺失一些数据的面板数据集。

unbiased estimator 无偏估计量：偏差等于零的估计量。

uncorrelated 不相关的：如果两个随机变量的相关关系是零，那么它们是不相关的。

underidentification 不足识别：当工具变量数小于内生回归因子数时。

unit root 单位根：指最大的根等于 1 的自回归。

unrestricted regression 无约束的回归：当计算经验规则 F 统计量时，这是备择假设下所应用的回归，因此，没有对系数加以约束以满足零假设。

VAR：请见*向量自回归*。

variance *方差*：一个随机变量与其均值之差的平方的期望值，Y 的方差用 σ_Y^2 表示。

vector autoregression 向量自回归：由 k 个方程所构成的 k 个时间序列变量的模型，每个方程对应着一个变量。其中，所有方程的回归因子是所有变量的滞后值。

volatility clustering 波动集聚：当一个时间序列变量在一些时期显示出一些高集聚的方差，而在另一些时期显示出一些低集聚的方差时。

weak instruments 弱工具变量：与内生回归因子有低相关关系的工具变量。

weighted least squares（WLS）加权最小二乘：当回归误差是异方差的，以及异方差的形式是已知的或能被估计时，可使用的一种替代 OLS 的方法。

学出版社的编辑高鹏先生和孙越女士，两位编辑严谨负责的工作态度使本书的翻译错误减少到了最低程度。

王庆石

2004.8